

基于多目标 PSO 混合优化的虚拟样本生成

王丹丹^{1,2,3} 汤健^{1,2,3} 夏恒^{1,2,3} 乔俊飞^{1,2,3}

摘要 受限于检测技术难度、高时间与经济成本等原因, 难测参数的软测量模型建模样本存在数量少、分布稀疏与不平衡等问题, 严重制约了数据驱动模型的泛化性能. 针对以上问题, 提出一种基于多目标粒子群优化 (Multi-objective particle swarm optimization, MOPSO) 混合优化的虚拟样本生成 (Virtual sample generation, VSG) 方法. 首先, 设计综合学习粒子群优化算法的种群表征机制, 使其能够同时编码用于连续变量和离散变量; 然后, 定义具有多阶段多目标特性的综合学习粒子群优化算法适应度函数, 使其能够在确保模型泛化性能的同时最小化虚拟样本数量; 最后, 提出面向虚拟样本生成的多目标混合优化任务以改进综合学习粒子群优化算法, 使其能够适应虚拟样本优选过程的变维特性并提高收敛速度. 同时, 首次借鉴度量学习提出用于评价虚拟样本质量的综合评价指标和分布相似指标. 利用基准数据集和真实工业数据集验证了所提方法的有效性和优越性.

关键词 小样本建模, 虚拟样本生成, 混合优化, 多目标粒子群优化, 分布相似度

引用格式 王丹丹, 汤健, 夏恒, 乔俊飞. 基于多目标 PSO 混合优化的虚拟样本生成. 自动化学报, 2024, 50(4): 790–811

DOI 10.16383/j.aas.c211091

Virtual Sample Generation Method Based on Hybrid Optimization With Multi-objective PSO

WANG Dan-Dan^{1,2,3} TANG Jian^{1,2,3} XIA Heng^{1,2,3} QIAO Jun-Fei^{1,2,3}

Abstract Due to the difficulty of detection technology, and high time and economic cost, the modeling samples of soft-sensing model with difficult parameters have some problems, such as small numbers, sparse distribution, and imbalance, which seriously restrict the generalization performance of data-driven models. To solve the above problems, a virtual sample generation (VSG) method based on multi-objective particle swarm optimization (MOPSO) hybrid optimization is proposed. First, the population representation mechanism of the integrated learning particle swarm optimization algorithm is designed, so that it can simultaneously encode the continuous and the discrete variables. Then, the fitness function of the integrated learning particle swarm optimization algorithm with multi-stage and multi-objective characteristics is defined to minimize the number of virtual samples while ensuring the generalization performance of the model. Finally, a multi-objective hybrid optimization task is generated for virtual samples to improve the integrated learning particle swarm optimization algorithm, so that it can adapt to the variable dimension characteristics of the virtual sample optimization process and improve the convergence speed. At the same time, the comprehensive evaluation index and distribution similarity index are proposed for evaluating the quality of virtual samples by referring to metric learning for the first time. In this paper, two benchmark datasets and an actual industrial dataset are used to verify the effectiveness and superiority of the proposed method.

Key words Small sample modeling, virtual sample generation (VSG), hybrid optimization, multi-objective particle swarm optimization (MOPSO), distribution similarity

Citation Wang Dan-Dan, Tang Jian, Xia Heng, Qiao Jun-Fei. Virtual sample generation method based on hybrid optimization with multi-objective PSO. *Acta Automatica Sinica*, 2024, 50(4): 790–811

收稿日期 2021-11-18 录用日期 2022-06-16

Manuscript received November 18, 2021; accepted June 16, 2022

国家自然科学基金 (62073006, 62173120, 62021003), 北京市自然科学基金资助项目 (4212032, 4192009), 科技创新 2030——“新一代人工智能”重大项目 (2021ZD0112301, 2021ZD0112302) 资助

Supported by National Natural Science Foundation of China (62073006, 62173120, 62021003), Beijing Natural Science Foundation (4212032, 4192009), and National Key Research and Development Program of China (2021ZD0112301, 2021ZD0112302)

本文责任编辑 刘艳军

Recommended by Associate Editor LIU Yan-Jun

1. 北京工业大学信息学部 北京 100124 2. 北京工业大学智慧

本文采用符号的含义见表 1.

实现复杂工业过程的智能控制和绿色生产需要对产品质量、能耗物耗、污染排放等难测参数 (如城

环保北京实验室 北京 100124 3. 北京工业大学智能感知与自主控制教育部工程研究中心 北京 100124

1. Faculty of Information Technology, Beijing University of Technology, Beijing 100124 2. Beijing Laboratory of Smart Environmental Protection, Beijing University of Technology, Beijing 100124 3. Engineering Research Center of Intelligent Perception and Autonomous Control, Ministry of Education, Beijing University of Technology, Beijing 100124

表 1 本文采用符号的含义
Table 1 The meaning of the symbols used in this article

序号	符号	含义
1	ρ_i	全局最优粒子选择指标
2	ρ_j	虚拟样本综合评价指标
3	η	数据分布相似度
4	$\mathbf{F}(\mathbf{z})$	多目标优化问题的目标函数集
5	$\mathbf{z}, \mathbf{z}_n^p(t+1)$	优化问题决策变量 (粒子的位置矢量), 表示第 $t+1$ 次迭代时, 粒子 p 的第 n 维位置值
6	$\mathbf{v}, \mathbf{v}_n^p(t+1)$	粒子的速度矢量, 表示第 $t+1$ 次迭代时, 粒子 p 的第 n 维速度值
7	$w_{inertia}$	粒子速度更新的惯性权重
8	$\mathbf{d}^p(t+1)$	第 $t+1$ 次迭代时, 粒子 p 的个体最优
9	E_n^p	粒子 p 的第 n 维的学习样例值
10	$N_{refresh}$	个体最优未更新阈值, 用于控制学习样例的更新
11	P_c^p	粒子 p 的学习概率, 用于控制学习样例的更新概率
12	$rank^p$	粒子 p 个体最优的适应度在种群中排名
13	K	RF 模型中决策树数量
14	L_F	RF 模型中切分特征数
15	θ_{leaf}	RF 模型中决策树的叶节点包含样本数量的阈值
16	F_{sel}^q	RF 模型中决策树的节点 q 最佳切分特征
17	s^q	RF 模型中决策树的节点 q 最佳分裂点取值
18	$f_{tree}^k(\cdot)$	RF 模型中第 k 个决策树模型
19	$f_{RF}(\cdot)$	RF 模型
20	\mathbf{z}_{para}	指导候选虚拟样本生成的参数决策变量
21	\mathbf{z}_{vss}	筛选候选虚拟样本选择决策变量
22	\mathbf{R}_{train}	原始小样本训练集
23	$\mathbf{x}_{vsg-min}, \mathbf{x}_{vsg-max}$	采用改进 MTD 进行扩展后的输入扩展域的上限和下限
24	$y_{vsg-min}, y_{vsg-max}$	采用改进 MTD 进行扩展后的输出扩展域的上限和下限
25	\mathbf{X}_{vs-g}	混合插值生成的虚拟样本输入
26	$\mathbf{X}_{equal}, \mathbf{X}_{rand}$	等间隔插值、随机插值生成的虚拟样本输入
27	$\mathbf{y}_{vs-g1}, \mathbf{y}_{vs-g2}$	基于虚拟样本输入, 结合 RF、RWNN 映射模型生成的虚拟样本输出
28	$\mathbf{R}_{vs-g1}^p, \mathbf{R}_{vs-g2}^p$	基于虚拟样本输入, 结合 RF、RWNN 映射模型生成的虚拟样本
29	\mathbf{R}_{vs-g}	生成的混合虚拟样本
30	\mathbf{R}_{vs-d}	对 \mathbf{R}_{vs-g} 进行删减后的候选虚拟样本
31	\mathbf{R}_{vs-s}	对候选虚拟样本进行选择后获得的虚拟样本
32	\mathbf{R}_{valid}	原始小样本验证集
33	\mathbf{R}_{vs}	最优虚拟样本
34	$f_{num}(\mathbf{z})$	多目标优化问题的目标之一, 筛选后的虚拟样本数量
35	$f_{mod}(\mathbf{z})$	多目标优化问题的目标之一, 筛选后的虚拟样本与原始训练集构建 RF 模型的性能指标
36	z_{MTD}	粒子的参数决策变量之一, 对应基于 MTD 方法的扩展率 γ_{extend}
37	z_{RF}^1	粒子的参数决策变量之一, 对应 RF 映射模型的切分特征数 L_F
38	z_{RF}^2	粒子的参数决策变量之一, 对应 RF 映射模型中决策树的中叶节点包含样本数量的阈值 θ_{leaf}
39	z_{RWNN}	粒子的参数决策变量之一, 对应 RWNN 映射模型的隐含层神经元数量 I
40	γ_{extend}	基于 MTD 方法的扩展率
41	I	RWNN 映射模型的隐含层神经元数量

表 1 本文采用符号的含义 (续表)

Table 1 The meaning of the symbols used in this article (continued table)

序号	符号	含义
42	$\mathbf{X}_{\text{train}}$	原始小样本训练集输入
43	$\mathbf{y}_{\text{train}}$	原始小样本训练集输出
44	y_{ave}	$\mathbf{y}_{\text{train}}$ 的均值
45	$\mathbf{y}_{\text{high}}, \mathbf{y}_{\text{low}}$	$\mathbf{X}_{\text{train}}$ 中大于/小于 y_{ave} 的输出集合
46	$y_{\text{max}}, y_{\text{min}}$	$\mathbf{y}_{\text{train}}$ 中最大值、最小值
47	$y_{\text{H-ave}}, y_{\text{L-ave}}$	$\mathbf{y}_{\text{high}}, \mathbf{y}_{\text{low}}$ 的均值
48	$N_{\text{equal}}, N_{\text{rand}}$	等间隔插值、随机插值倍数
49	\mathbf{W}, \mathbf{b}	RWNN 模型输入层与隐含层间神经元的连接权重与偏置
50	\mathbf{H}_{ori}	RWNN 模型隐含层输出矩阵
51	β	RWNN 模型隐含层与输出层神经元的连接权重
52	$N_{\text{vs-g}}, N_{\text{vs-d}}, N_{\text{vs-s}}$	生成、候选、选择后虚拟样本的数量
53	θ_{select}	虚拟样本的选择阈值
54	$\tilde{\mathbf{z}}_{\text{vss}}$	对 \mathbf{z}_{vss} 进行变维度处理后获得
55	F	使用虚拟样本集 $\mathbf{R}_{\text{vs-s}}$ 的建模性能指标
56	\mathbf{R}_{mix}	原始训练集 $\mathbf{R}_{\text{train}}$ 与 $\mathbf{R}_{\text{vs-s}}$ 的混合样本集
57	P_{num}	种群中粒子数量
58	N_{iter}	种群迭代次数
59	\mathbf{A}	种群的外部档案, 保存非支配解

市固废焚烧 (Municipal solid waste incineration, MSWI) 过程中的有机污染物二噁英 (Dioxin, DXN) 的排放浓度^[1] 等进行实时检测^[2]. MSWI 是目前世界范围内应用最为广泛的城市固废无害化、减量化和资源化处理手段^[3-4] 以及国家“十四五”规划鼓励推行技术, 该过程中被严格限制排放的 DXN 被称作“世纪之毒”^[5]. 以实时、准确、低成本方式实现 DXN 的检测是降低其排放控制的关键技术之一, 也是目前业界亟待解决的难题^[6]. 因工业过程长期在稳态模式下运行, 这使得现场采集的数据所对应的工况极为相似, 通过实验设计方式或突发工况情景获取非稳态模式过程数据、异常数据甚至故障数据的风险很高或不被允许, 进而导致有效建模样本数据稀少且分布不均衡^[7-8]. 另外, 诸如选矿磨矿^[9]、柔性制造^[10] 和化工生产^[11] 等工业过程, 由于实时进行难测参数真值检测的技术难度大、离线化验的时间与经济成本高等原因, 使得工业过程难测参数建模面临着“大数据、小样本”问题^[12]. 目前, 通过虚拟样本生成 (Virtual sample generation, VSG) 技术扩充建模样本数量已成为解决上述小样本问题的有效手段之一, 也是目前学术界的研究的难点和热点^[9].

由模式识别领域首次提出的 VSG 技术通过扩增原始建模样本的方法, 解决面向分类的小样本问题^[13], 其本质是通过撷取小样本间的缺失信息生成

适当数量的虚拟样本^[14], Niyogi 等^[15] 从数学上证明了 VSG 等效于正则化策略. 目前, VSG 技术已被成功地应用于癌症识别^[16]、可靠性分析^[17]、机械振动信号建模^[9] 等领域, 其在图像识别领域的应用尤为广泛^[18-21]. 主要策略是结合先验知识, 通过几何变换等操作生成虚拟图像. 针对复杂工业过程, 只有具有长期运行经验的领域专家才能抽象出明确的先验知识, 但也存在一定的主观性和随意性. 针对先验知识无法获取或提取难度大的问题, VSG 的研究开始聚焦于如何从已知样本中汲取知识以生成虚拟样本. Li 等^[22] 为解决制造系统早期样本较少问题, 提出基于区间核密度估计的 VSG, 核心是根据小样本数据估计总体分布后再生成虚拟样本. 进一步, Li 等^[23] 和 Lin 等^[24] 分别提出了基于双参数威布尔分布估计和多模态分布估计的 VSG. 针对上述研究存在小样本分布不均衡情况下估计偏差较大的问题, Li 等^[16] 提出基于模糊理论信息扩散准则的整体趋势扩散 (Mega-trend-diffusion, MTD) 技术, 本质是通过数据分布趋势扩展样本空间, 并在扩展域内生成虚拟样本. 上述 VSG 研究主要面向分类问题, 特点在于仅需要为不同类别生成虚拟样本的输入即可; 相对于本文所面对的回归建模问题, 还需要考虑如何为合理的虚拟样本输入生成精准的虚拟输出. 因此, 面向回归的 VSG 的研究难度较大, 这也是相关

文献较少的原因之一。

为使得虚拟样本输入能够均衡地填补真实小样本间的信息间隙, Zhu 等^[11] 先利用距离准则识别信息空隙区域, 再进行 Kriging 插值; Zhang 等^[25] 先采用流形学习 Isomap 识别样本稀疏区域, 再进行插值; Chen 等^[26] 先采用查询策略获取稀疏区域, 再进行插值. 进一步, 同时考虑虚拟样本的输入和输出, Li 等^[27] 先基于树的趋势扩散技术进行区域扩展后, 再依据启发式机制同时生成输入与输出; Zhu 等^[28] 先依据多分布趋势扩散技术生成虚拟样本输入, 再通过小样本映射模型生成输出; He 等^[29] 和朱宝等^[30] 基于神经网络模型隐含层插值和缩放方式, 同时生成非线性输入与输出; Qiao 等^[31] 结合改进 MTD 技术与隐含层插值生成输入与输出. 此外, 针对物理含义清晰的工业过程实验数据, Tang 等^[32] 通过线性插值生成虚拟样本输入后, 再依据多个映射模型融合生成相应输出. 针对虚拟样本输入输出难以有效获得的问题, Li 等^[33] 先通过 MTD 进行域扩展再采用遗传算法 (Genetic algorithm, GA) 生成优化虚拟样本, Chen 等^[34] 采用粒子群优化 (Particle swarm optimization, PSO) 算法生成虚拟样本. 上述算法的优点是同时考虑了数据属性间的相互影响, 但未予考虑所虚拟样本间的多样性和映射模型超参数对虚拟样本的影响.

总之, 为生成更为合理的虚拟样本, 已经存在诸多 VSG 方法. 考虑到虚拟样本与实际数据间存在的偏差, 这些不同方法所生成的虚拟样本间也必然存在着冗余性与互补性. 对此, 汤健等^[35] 提出面向已经生成的虚拟样本的优化选择策略, 虽然采用的用于获取虚拟样本输出的随机神经网络 (Random weight neural network, RWNN) 映射模型具有结构简单、计算复杂度低、能够进行隐含层插值等特点, 但其固有的随机性使得所生成的虚拟样本输出精度难以保证. 随机森林 (Random forest, RF) 对于多数数据集均具有良好的表现, 能够处理具有离散、连续、高维等特性的数据^[36]. 显然, RF 作为生成虚拟样本输出的映射模型可以提高虚拟样本的质量. 此外, 由于映射模型的超参数取值影响虚拟样本的质量, 因此在生成虚拟样本的过程中, 对强关联性的超参数进行优化也是提高 VSG 的一个改进方向. 显然, 对映射模型的超参数和虚拟样本的选择进行同时优化属于连续变量和离散变量的混合优化问题, 这不仅需要确保超参数的优化过程不会提前收敛至局部最优, 也需要在进行大量虚拟样本优化选择时, 具有较好的收敛速度. 研究表明, 综合学习粒子群优化 (Comprehensive learning particle swarm

optimization, CLPSO) 算法依据所有其他粒子的历史最佳信息进行粒子更新, 能够保持种群多样性且防止过早收敛^[37]. 此外, 笔者认为, 筛除冗余虚拟样本的关键在于如何对虚拟样本进行合理评价, 但目前对该问题的研究还不够深入. 另外, 由于虚拟样本引入的预测误差存在积累效应, 这使得虚拟样本的数量会影响建模性能; 但是, 以往研究主要通过实验确定虚拟样本最佳数量^[38]. 林越等^[39] 基于信息熵理论推导得到虚拟样本的最佳数量, 但是实际上虚拟样本的最佳数量往往与建模数据质量具有较大的相关性. 显然, 有必要通过多目标优化策略实现对虚拟样本数量和质量的综合均衡.

综上所述, 面向工业过程回归建模的 VSG 研究存在以下难点: 1) 针对原始小样本的分布稀疏与不均衡特性, 如何基于原始小样本探究实际数据的分布空间, 均衡地生成虚拟样本输入; 2) 如何通过映射模型为虚拟输入生成合理的虚拟输出, 获得大量高质量具有冗余与互补特性的虚拟样本; 3) 如何筛选出有效的高质量虚拟样本并确定其最佳数量; 4) 如何对虚拟样本进行量化评价以支撑其筛选策略.

针对上述亟待解决的难点, 结合笔者已有研究成果, 本文提出一种基于多目标 PSO 混合优化的虚拟样本生成策略, 用于优化虚拟样本的生成与选择过程, 包括面向混合优化的粒子设计、面向 VSG 的适应度函数设计和面向 VSG 的多目标混合优化. 本文首次提出将 VSG 问题描述为多目标混合优化任务, 并首次采用度量学习的指标对虚拟样本的质量进行评价. 通过基准数据集和实际工业数据集实验, 验证了本文 VSG 方法的合理性和有效性.

1 相关知识

1.1 小样本数据回归建模

对工业过程难测参数进行软测量建模 (即通过机器学习方法构建易测过程变量与难测参数间的映射模型) 是目前业界的常用检测手段^[40]. 例如在 MSWI 过程中, DXN 排放浓度通常采用离线直接检测法和在线间接检测法, 但以上 2 种方法均存在价格昂贵、时间滞后等局限性, 难以支撑以降低污染排放浓度为目标的实时运行优化^[1]. 构建基于 MSWI 过程易测变量的 DXN 排放浓度软测量模型, 虽然能够克服以上问题, 但 DXN 排放浓度的有标记真值样本 (真输入-真输出) 具有高维、稀缺的特性^[41]. 综上所述, 对工业过程难测参数进行建模, 往往需要在解决建模样本维度高、数量少、分布稀疏与不

平衡等问题后,才能构建得到具有高精度、强鲁棒性能的软测量模型。

理论上,数据驱动建模通常用于建模样本足够丰富且真值获取成本相对较低的场景^[42].统计学科认为,建模样本数量应该大于等于输入特征维数或大于等于 30^[43].众多研究学者指出,小样本是指有效样本数量少于 30 (或 50) 或样本数量少于输入特征维数的 k 倍 (k 取 2、5、10)^[9, 44-46].可见,小样本问题不能简单理解为样本绝对数量较少,而是与输入特征维数有关的相对概念,其本质是样本中所包含的建模所需特征信息不足.另外,小样本数据也存在分布稀疏与不平衡等特性^[47].因此,基于小样本构建的软测量模型往往具有片面性和偏差性,难以实现难测参数的有效预测.目前,已有多种机器学习方法用于改善小样本数据的建模性能,包括支持向量机^[47-48]、灰色模型^[49]、核回归^[50]和贝叶斯网络^[51-52]等.在样本数量稀缺及分布不平衡的情况下,上述算法也难以进一步提高软测量模型的预测精度.因此,需要从新的视角解决工业小样本数据的回归建模问题。

1.2 虚拟样本生成

1.2.1 虚拟样本定义

1992年, Poggio等^[13]首次提出 VSG 方法用于人脸识别问题.进一步,文献^[53]给出虚拟样本定义如下。

定义 1. 对于给定的训练样本 (\mathbf{x}, \mathbf{y}) , 通过变换 T 得到的样本 $(T\mathbf{x}, y_T(\mathbf{x}))$ 也是合理样本, 那么新得到的样本 $(T\mathbf{x}, y_T(\mathbf{x}))$ 就是通过变换 T 生成的虚拟样本:

$$\left. \begin{array}{l} (\mathbf{x}, \mathbf{y}) \\ Know \end{array} \right\} \xrightarrow{(T, y_T(\cdot))} (T\mathbf{x}, y_T(\mathbf{x})) \quad (1)$$

式中, 变换 $(T, y_T(\cdot))$ 即为领域先验知识 $Know$. 通常, 先验知识包括: 1) 直接从问题中提取物理含义明确的知识; 2) 从小样本中获取先验知识; 3) 在学习算法中嵌入先验知识等。

1.2.2 虚拟样本内涵

VSG 的本质是依据小样本数据生成尽可能符合真实数据分布的虚拟样本. 虚拟样本与真实样本间的关系如图 1 所示. 图 1 展示了虚拟样本、真实样本、小样本空间、虚拟样本空间、实际数据空间之间的关系。

由图 1 可知, 小样本存在如下问题: 1) 小样本未能全面覆盖实际数据空间, 存在信息空白区域; 2) 小样本间存在信息间隙; 3) 小样本未能在实际数

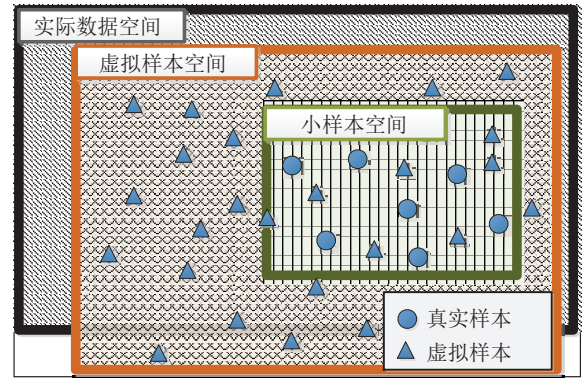


图 1 虚拟样本与真实样本间的关系

Fig.1 Relationship between virtual samples and real samples

据空间中均匀分布. 因此, 小样本空间只能片面反映实际数据空间. 众多学者研究 VSG 的目标是使虚拟样本空间能尽可能地贴近实际数据空间. 但无论采用哪种 VSG, 必然会生成某些不符合实际数据特征和分布的虚拟样本 (如图 1 下部所示的实际数据空间之外的虚拟样本), 其不仅不利于模型的训练, 还会导致模型泛化性能变差。

显然, 针对虚拟样本质量的评判问题, 需要提出更加合理的评价指标和筛选机制。

1.2.3 回归建模 VSG

面向工业过程回归建模的 VSG 问题比分类领域的难度更大, 如何生成虚拟样本的输入和输出是主要焦点. 生成的虚拟样本输入应具有的特征包括: 1) 能够贴近实际数据分布; 2) 可填补小样本的信息间隙或空白; 3) 可缓解小样本分布的不均衡性. 生成虚拟样本输出的方法是先构建基于小样本的映射模型再预测输出, 当平均绝对百分比误差不超过 10% 时, 可用于生成虚拟输出. 虽然通过调整模型参数可达到上述要求, 但由于映射模型构建方法固有的差异性, 采用相同虚拟输入映射得到的输出在稳定性和扩展性上存在较大差异. 为得到更为合理的虚拟样本输出, 映射模型应该具有较好的数据适应性. 另外, 为消除所生成虚拟样本间存在的冗余性, 汤健等^[35]采用 PSO 算法对虚拟样本进行了优化选择. 如何确定虚拟样本数量和评价其质量, 还是开放问题。

综上, 有必要从同时优化虚拟样本质量和模型泛化性能的视角求解 VSG 问题。

1.3 多目标粒子群优化

1.3.1 多目标优化问题

通常, 多目标优化问题 (Multi-objective op-

timization problem, MOP) 被转化为最小化优化问题进行研究, 其描述为:

$$\begin{aligned} \min \mathbf{F}(\mathbf{z}) &= (f_1(\mathbf{z}), f_2(\mathbf{z}), \dots, f_m(\mathbf{z})) \\ \text{s.t. } \mathbf{z} &\in \Omega \end{aligned} \quad (2)$$

式中, $\mathbf{z} = (z_1, z_2, \dots, z_n)$ 为决策变量, Ω 表示可行搜索域, $\mathbf{F}(\mathbf{z}) : \Omega \rightarrow \mathbf{S}$ 是由 m 个实值函数组成的优化目标, \mathbf{S} 表示目标空间.

设 $\mathbf{a}, \mathbf{b} \in \Omega$ 为式 (2) 定义的 MOP 的 2 个可行解. 当且仅当对于任意 $i \in \{1, 2, \dots, m\}$, 都有 $f_i(\mathbf{a}) \leq f_i(\mathbf{b})$ 且至少有一个 $j \in \{1, 2, \dots, m\}$, 使得 $f_j(\mathbf{a}) < f_j(\mathbf{b})$ 时, \mathbf{a} 支配 \mathbf{b} , 记作 $\mathbf{a} \prec \mathbf{b}$. 如果不存在 $\mathbf{a} \in \Omega$, 使得 $\mathbf{F}(\mathbf{a})$ 支配 $\mathbf{F}(\mathbf{a}^*)$, 那么 \mathbf{a}^* 是式 (2) 定义的 MOP 的一个 Pareto 最优解, $\mathbf{F}(\mathbf{a}^*)$ 为 Pareto 最优 (目标) 矢量. 显然, Pareto 最优解中任何一个目标性能的提升必然导致至少一个其他目标性能的下降. 通常, 所有 Pareto 最优解的集合称为 Pareto 最优解集 (Pareto optimal solution set, POS), 所有 Pareto 最优目标矢量的集合称为 Pareto 最优前沿.

求解多目标优化问题的常用进化算法包括遗传算法、差分进化 (Differential evolution, DE) 算法和 PSO 等. GA 算法通过选择、交叉和变异等操作产生新解, 适用于离散型的优化问题, 其运行时间随种群规模指数级增长. DE 算法随机选择 3 个与自身不同的个体生成新个体, 通过实数编码对可行域进行搜索, 其超参数对算法性能影响较小, 收敛性能好, 但针对混合优化 DE 算法的研究很少. 标准 PSO 算法是模拟鸟群捕食行为的智能优化算法, 其原理是通过种群中个体间的相互协作和信息共享寻找最优解, 其粒子跟随全局最优与个体最优位置进行移动, 虽然搜索空间连续, 但也可求解特征选择等离散问题. 标准 PSO 算法容易陷入局部最优解, 且当全局最优与个体最优矛盾时会造成算力的浪费.

1.3.2 综合学习 PSO 描述

相对于标准 PSO 算法, CLPSO 算法对粒子速度的更新策略进行了改进, 提高算法的全局搜索能力, 其粒子速度 \mathbf{v}^p 与位置 \mathbf{z}^p 的更新公式如下:

$$\mathbf{v}_n^p(t+1) = w_{\text{inertia}}(t) \cdot \mathbf{v}_n^p(t) + c \cdot r_n^p \cdot (E_n^p(t) - \mathbf{z}_n^p(t)) \quad (3)$$

$$\mathbf{z}_n^p(t+1) = \mathbf{z}_n^p(t) + \mathbf{v}_n^p(t+1) \quad (4)$$

式中, w_{inertia} 是影响粒子搜索步长的惯性权重, c 为学习因子, r_n^p 服从 $[0, 1]$ 间的均匀分布, E_n^p 为粒子 p 第 n 维的学习样例.

由式 (3) 可知, 粒子速度的更新不再受个体最优与全局最优的综合影响, 而是学习所有粒子的个体最优, 其更新公式如下:

$$\mathbf{d}^p(t+1) = \begin{cases} \mathbf{z}^p(t+1), & \mathbf{z}^p(t+1) \prec \mathbf{d}^p(t) \\ \mathbf{d}^p(t), & \text{其他} \end{cases} \quad (5)$$

式中, $\mathbf{d}^p = (d_1^p, d_2^p, \dots, d_n^p)$ 表示粒子 p 的个体最优.

CLPSO 为每个粒子均维持一个样例池, 粒子各个维度学习其相应的样例. 显然, 该策略能够保持种群多样性, 有效缓解标准 PSO 提前收敛的问题. 若粒子个体最优迭代 N_{refresh} 次后仍未能更新, 则更新其学习样例池. 策略为: 设定粒子各维度学习样例的更新概率为 P_c^p , 更新时, 首先任意选择种群中 2 个粒子, 然后对比 2 个粒子的个体最优, 竞争选择较好的个体最优作为新学习样例, 可表示为:

$$E_n^p = \left\{ \mathbf{d}_n^{p'} \mid \min_{p'}(f(\mathbf{d}^{p'})), \quad p' = p_{\text{rand1}}, p_{\text{rand2}} \right\} \quad (6)$$

P_c^p 为粒子 p 的学习概率, 更新如下:

$$P_c^p = 0.05 + 0.45 \frac{e^{\frac{10(\text{rank}^p - 1)}{N-1}} - 1}{e^{10} - 1} \quad (7)$$

式中, rank^p 表示粒子个体最优的适应度排名, 随着粒子的排序 rank^p 递增, 其学习概率随之增大, 即学习样例的更新概率从 5% 逐渐增大到 50%.

2 基于 MOPSO 混合优化的 VSG 策略

综上, 本文提出的对虚拟样本生成与选择过程进行多目标混合优化方法, 称为 MoHo-VSG. 首先, 对相关超参数进行优化以获取候选虚拟样本; 然后, 对后者进行优化选择以获得最优虚拟样本. 基于多目标粒子群优化 (Multi-objective particle swarm optimization, MOPSO) 混合优化的虚拟样本生成策略如图 2 所示. 该方法由面向混合优化的粒子设计模块、面向 VSG 的适应度函数设计模块和面向 VSG 的多目标混合优化模块组成. 图 2 中, \mathbf{z}^p 表示优化问题的决策变量, 即 PSO 中粒子的位置, 包括参数决策变量 $\mathbf{z}_{\text{para}}^p$ 和样本选择决策变量 $\mathbf{z}_{\text{vss}}^p$; $\mathbf{R}_{\text{train}}$ 和 $\mathbf{R}_{\text{valid}}$ 分别表示由原始小样本划分得到的训练集和验证集; $\mathbf{x}_{\text{vsg-max}}^p$ 和 $\mathbf{x}_{\text{vsg-min}}^p$ 分别表示采用改进 MTD 方法对域进行扩展得到输入扩展域的上限和下限, $y_{\text{vsg-max}}^p$ 和 $y_{\text{vsg-min}}^p$ 为相应的输出扩展域的上限和下限; $\mathbf{X}_{\text{vs-g}}^p$ 表示在扩展域的上/下限中, 通过混合插值方法生成的虚拟样本输入; $\mathbf{R}_{\text{vs-g1}}^p$ 和 $\mathbf{R}_{\text{vs-g2}}^p$ 分别表示通过 RF 和 RWNN 映射模型获得的虚拟样本集; $\mathbf{R}_{\text{vs-d}}^p$ 为 $\mathbf{R}_{\text{vs-g1}}^p$ 和 $\mathbf{R}_{\text{vs-g2}}^p$ 混合后, 依据输出扩展域的

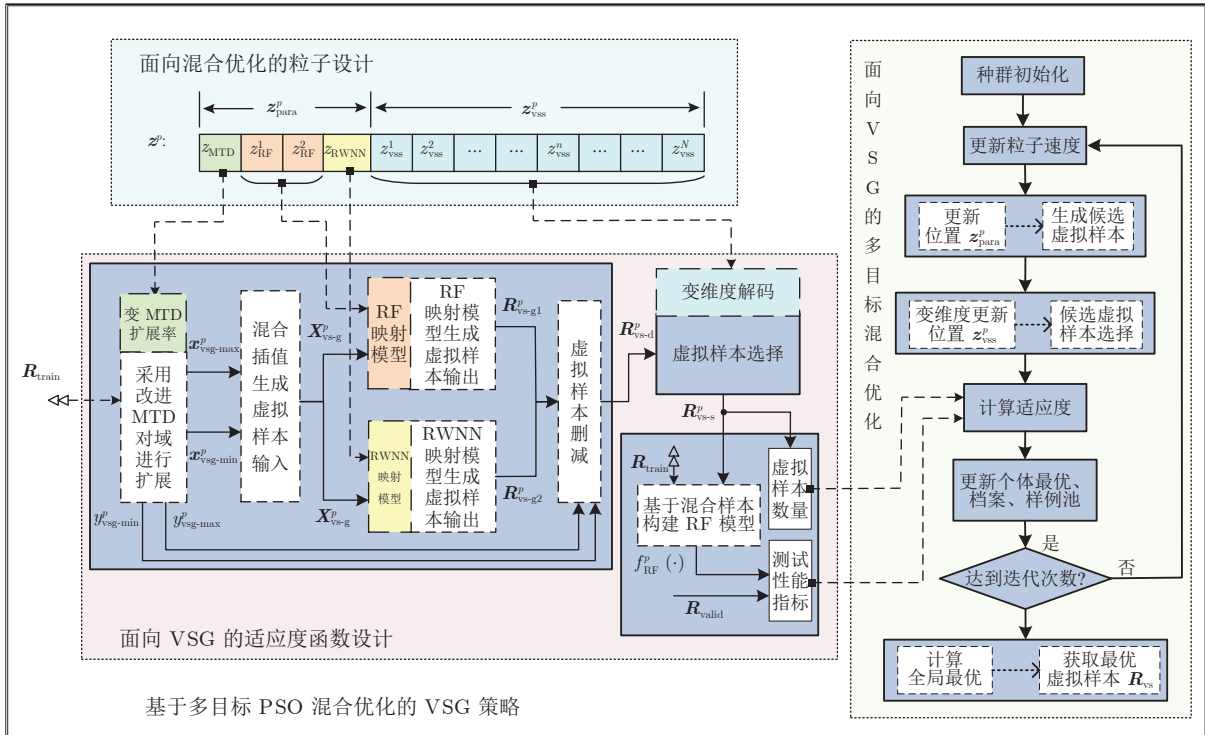


图 2 基于 MOPSO 混合优化的 VSG 策略
Fig.2 VSG based on hybrid optimization with MOPSO

上/下限删减获得的候选虚拟样本集; R_{vs-s}^p 表示 R_{vs-d}^p 经粒子选择后获得的优选虚拟样本; $f_{RF}^p(\cdot)$ 为由混合样本构建的 RF 模型; R_{vs} 为根据全局最优获得的最优虚拟样本. 主要模块功能如下:

1) 面向混合优化的粒子设计

将决策变量分为参数决策和样本选择决策变量 2 个部分, 前者为指导候选虚拟样本生成的连续变量, 后者为筛选候选虚拟样本的高维离散变量, 通过粒子设计实现混合优化的策略.

2) 面向 VSG 的适应度函数设计

分为生成候选虚拟样本、候选虚拟样本选择和虚拟样本评价指标计算共 3 个阶段计算适应度, 评价指标包括虚拟样本数量和混合样本构建模型在验证集上的预测性能.

3) 面向 VSG 的多目标混合优化

改进 CLPSO 算法, 以适应 VSG 过程的变维度特性. 在达到最大迭代次数和确定全局最优后, 获得最优虚拟样本集.

3 基于 MOPSO 混合优化的 VSG 实现

本文采用多目标优化的目的是, 在确保虚拟样本达到最优建模效果的前提下, 尽可能地减少其数量. 相应地, 本文的优化目标可描述为:

$$\begin{aligned} \min \mathbf{F}(\mathbf{z}) &= (f_{\text{num}}(\mathbf{z}), f_{\text{mod}}(\mathbf{z})) \\ \text{s.t.} \\ \mathbf{z} &\in \Omega \end{aligned} \tag{8}$$

式中, 决策矢量 \mathbf{z} 指导虚拟样本的生成和筛选, $f_{\text{num}}(\mathbf{z})$ 表示筛选后虚拟样本的数量, $f_{\text{mod}}(\mathbf{z})$ 表示筛选后由虚拟样本与训练集混合后构建的 RF 模型性能指标.

3.1 面向混合优化的粒子设计

基于混合优化策略对粒子进行设计如图 3 所示. 本文设计的粒子记为 $\mathbf{z}^p = \{\mathbf{z}_{\text{para}}^p, \mathbf{z}_{\text{vss}}^p\}$, 其中 $\mathbf{z}_{\text{para}}^p = \{z_{\text{MTD}}^1, z_{\text{RF}}^1, z_{\text{RF}}^2, z_{\text{RWNN}}\}$ 为参数决策变量, 用于虚拟样本生成过程中映射超参数的优化; $\mathbf{z}_{\text{vss}}^p = \{z_{\text{vss}}^1, z_{\text{vss}}^2, \dots, z_{\text{vss}}^n, \dots, z_{\text{vss}}^N\}$ 为样本选择决策变量, 用于候选虚拟样本的优化选择. 由图 3 可知:

1) 每个粒子都包含位置 \mathbf{z}^p 、速度 \mathbf{v}^p 、学习样例 \mathbf{E}^p 、适应度 $f_{\text{num}}(\mathbf{z}^p)$ 和 $f_{\text{mod}}(\mathbf{z}^p)$ 、个体最优排序 rank^p 和学习概率 P_c^p 等属性, 其中粒子的个体最优排序 rank^p 确定了学习概率 P_c^p , 进而影响学习样例 \mathbf{E}^p 的更新; \mathbf{E}^p 指导粒子速度 \mathbf{v}^p 的搜索方向和步长, 进而决定了粒子的位置更新. 根据粒子位置 \mathbf{z}^p 计算适应度 $f_{\text{num}}(\mathbf{z}^p)$ 和 $f_{\text{mod}}(\mathbf{z}^p)$, 通过适应度更新新粒子的个体最优及其排序.

2) 用于优化虚拟样本生成过程的参数决策变量 $\mathbf{z}_{\text{para}}^p = \{z_{\text{MTD}}, z_{\text{RF}}^1, z_{\text{RF}}^2, z_{\text{RWNN}}\}$ 包含 4 个参数, 其

z^p :	z_{MTD}^1	z_{RF}^1	z_{RF}^2	z_{RWNN}	z_{vss}^1	z_{vss}^2	z_{vss}^3	...	z_{vss}^n	z_{vss}^{n+1}	...	z_{vss}^N
位置 z^p :	0.56	37.00	13.00	4.00	0.32	0.03	0.67	...	0.98	0.43	...	0.55
速度 v^p :	0.02	-4.30	1.70	2.10	0.34	0.40	-0.54	...	0.02	-0.18	...	0.49
样例 E^p :	0.68	29.00	15.00	14.00	0.45	0.37	0.42	...	0.99	0.12	...	0.67
$f_{\text{min}}(z^p) = 78.00$ $f_{\text{mod}}(z^p) = 13.53$ $\text{rank}^p = 26.00$ $P_e^p = 0.16$												

图 3 基于混合优化策略的粒子设计

Fig.3 Particle design based on hybrid optimization strategy

中 z_{MTD} 为基于 MTD 方法的扩展率 γ_{extend} , 可依据不同建模数据的分布情况优化其取值, 从而获得更符合真实数据的扩展空间, 并在此空间中生成虚拟样本输入; z_{RF}^1 和 z_{RF}^2 分别表示 RF 映射模型的切分特征数 L_F 和决策树叶节点包含样本数量的阈值 θ_{leaf} , 通过优化 L_F 和 θ_{leaf} 2 个参数, 以构建更能反映真实数据特征的映射模型, 从而生成更精确的虚拟样本输出; z_{RWNN} 表示 RWNN 映射模型的隐含层神经元数量 I , 优化该参数的目的同 RF 映射模型.

3) 用于虚拟样本优化选择的样本选择决策变量 $z_{\text{vss}}^p = \{z_{\text{vss}}^1, z_{\text{vss}}^2, \dots, z_{\text{vss}}^n, \dots, z_{\text{vss}}^N\}$ 的维数与待选择的虚拟样本数量一致, 其中 $z_{\text{vss}}^n \in [0, 1]^{\mathbf{R}}$. 在优化过程中, z_{vss}^p 通过编解码的方式对虚拟样本进行选择. 另外, 由于迭代过程中候选虚拟样本的数不固定, 相应 z_{vss}^p 的维度也需要进行变维度处理.

3.2 面向 VSG 的适应度函数设计

适应度函数的设计即根据粒子的位置 z^p 计算式 (8) 所定义优化目标的过程. 本文的目标是对虚拟样本的生成和选择过程进行混合优化, 即通过粒子位置 z^p 指导虚拟样本的生成和选择, 将虚拟样本的数量和质量同时作为粒子的适应度. 因此, 本文面向 VSG 的适应度函数设计包含参数决策变量指导候选虚拟样本生成、样本选择决策变量对候选虚拟样本进行选择、虚拟样本评价指标计算 3 个部分.

3.2.1 生成候选虚拟样本

参数决策变量 $z_{\text{para}}^p = \{z_{\text{MTD}}, z_{\text{RF}}^1, z_{\text{RF}}^2, z_{\text{RWNN}}\}$ 与 MTD、RF 和 RWNN 超参数之间的关系如下:

$$\begin{cases} \gamma_{\text{extend}} = z_{\text{MTD}} \\ L_F = z_{\text{RF}}^1 \\ \theta_{\text{leaf}} = z_{\text{RF}}^2 \\ I = z_{\text{RWNN}} \end{cases} \quad (9)$$

生成候选虚拟样本的过程为: 首先, 基于扩展率 γ_{extend} , 对原始样本空间进行基于 MTD 的扩展, 在原始域和扩展域中, 通过混合插值生成虚拟样本输入; 然后, 基于 RF 建模参数 L_F 和 θ_{leaf} 构建 RF 映射模型, 基于 RWNN 建模参数 I 构建 RWNN 映射模型生成对应虚拟样本输出的输出值; 最后, 对生成的虚拟样本进行混合和删减, 以获得候选虚拟样本.

3.2.1.1 生成虚拟样本输入

基于扩展率 γ_{extend} , 采用改进 MTD 分别对原始训练集 $\mathbf{R}_{\text{train}} = \{\mathbf{X}_{\text{train}}, \mathbf{y}_{\text{train}}\} \in \mathbf{R}^{N \times L}$ 的输入和输出空间进行扩展.

1) 首先, 对输出域进行扩展. 计算 $\mathbf{y}_{\text{train}} = \{y_n\}_{n=1}^N$ 的均值 y_{ave} , 并由此将 $\mathbf{y}_{\text{train}}$ 分为大于均值的 \mathbf{y}_{high} 和小于均值的 \mathbf{y}_{low} 两部分后, 再计算 $\mathbf{y}_{\text{train}}$ 的最大值 y_{max} 和最小值 y_{min} , 作为扩展空间; 然后, 分别计算 \mathbf{y}_{high} 的均值 $y_{\text{H-ave}}$ 和 \mathbf{y}_{low} 的均值 $y_{\text{L-ave}}$; 最后, 计算获得输出扩展域的上限 $y_{\text{vsg-max}}$ 和下限 $y_{\text{vsg-min}}$ [35]. 以相同方式对样本输入空间进行扩展, 获得其扩展上限 $x_{\text{vsg-max}}$ 和下限 $x_{\text{vsg-min}}$.

2) 在样本输入扩展空间中, 进行等间隔插值和随机插值, 以生成虚拟样本输入.

首先, 分别在小样本空间和扩展空间进行 N_{equal} 倍的等间隔插值 [8], 获得等间隔插值虚拟样本输入, 记为 $\mathbf{X}_{\text{equal}}$.

然后, 在输入扩展空间进行随机插值, 获得随机插值虚拟样本输入, 记为 \mathbf{X}_{rand} :

$$\mathbf{x}_{\text{rand}}^n = x_{\text{vsg-min}} + \text{rand}^L \cdot (x_{\text{vsg-max}} - x_{\text{vsg-min}}) \quad (10)$$

$$\mathbf{X}_{\text{rand}} = \left\{ \mathbf{x}_{\text{rand}}^1, \mathbf{x}_{\text{rand}}^2, \dots, \mathbf{x}_{\text{rand}}^n, \dots, \mathbf{x}_{\text{rand}}^{N \cdot N_{\text{rand}}} \right\} \quad (11)$$

式中, N_{rand} 表示随机插值倍数, rand^L 表示第 n 个样本对应的随机值.

3) 将等间隔插值与随机插值获得的虚拟样本输入混合, 得到虚拟样本输入, 记为 $\mathbf{X}_{vs-g} = \{\mathbf{X}_{\text{equal}}; \mathbf{X}_{\text{rand}}\}$.

3.2.1.2 生成虚拟样本输出

为获得丰富的虚拟样本, 本文采用 2 个映射模型生成虚拟样本的输出, 其中 RF 和 RWNN 映射模型分别可获得稳定性较高和随机性较强的输出.

1) RF 映射模型的输出

基于参数 L_F 和 θ_{leaf} , 使用原始训练集 $\mathbf{R}_{\text{train}} = \{\mathbf{X}_{\text{train}}, \mathbf{y}_{\text{train}}\} \in \mathbf{R}^{N \times L}$ 构建 RF 映射模型.

首先, 对 $\mathbf{R}_{\text{train}}$ 进行有放回的 N 次随机采样, 从中随机切分出 L_F 个特征, 获得训练集 $\{\mathbf{R}_{\text{train}}^k\}_{N \times L_F}$, 由此构建第 k 个决策树模型 $f_{\text{tree}}^k(\cdot)$, 设定限制决策树继续分裂的最小样本数量为 θ_{leaf} . 对于回归问题, 决策树计算节点 q 的最佳切分特征 F_{sel}^q 和分裂点取值 s^q 的获取过程可表示为如下所示的优化问题^[36]:

$$(F_{\text{sel}}^q, s^q) = \arg \min \left(\sum_{i=1}^{N_{\text{left}}} (y_{\text{left}}^i - \bar{y}_{\text{left}})^2 + \sum_{i=1}^{N_{\text{right}}} (y_{\text{right}}^i - \bar{y}_{\text{right}})^2 \right) \quad (12)$$

式中, 决策树中各节点将样本分为左/右两部分. y_{left}^i 和 y_{right}^i 分别表示左/右分支上的样本输出, \bar{y}_{left} 和 \bar{y}_{right} 分别表示左/右分支上的样本输出均值, N_{left} 和 N_{right} 分别表示左/右分支上的样本数量.

重复上述过程, 构建得到 K 个决策树. 对上述全部决策树进行集成, 得到最终映射模型, 具体建模过程详见算法 1, 其中 θ_{leaf} 表示叶节点包含样本数量的阈值.

算法 1. RF 算法伪代码

- 1) 利用 Bootstrap 和随机子空间法对训练集 D 进行样本和特征的随机采样, 获得 K 个子训练集 $\{D_1, D_2, \dots, D_K\}$;
- 2) For $k=1$ to K ;
- 3) 根据式 (12), 遍历寻找最佳切分特征 F_{sel}^q 和切分点 s^q ;
- 4) 根据 F_{sel}^q 和 s^q 生成节点, 将输入特征空间分为左/右 2 个区域;
- 5) if 节点包含样本数量大于等于 θ_{leaf} ;
- 6) 重复步骤 3) ~ 4), 不断生成新的节点, 直到新节点包含样本数量小于 θ_{leaf} ;
- 7) End if;
- 8) 第 k 个回归树 $f_{\text{tree}}^k(\cdot)$ 构建完成;
- 9) End for;
- 10) 计算输出.

然后, 基于 RF 映射模型所获得的虚拟样本输出为:

$$\mathbf{y}_{vs-g1} = \frac{1}{K} \sum_{k=1}^K f_{\text{tree}}^k(\mathbf{X}_{vs-g}) \quad (13)$$

最后, 获得虚拟样本集, 记为 $\mathbf{R}_{vs-g1} = \{\mathbf{X}_{vs-g}, \mathbf{y}_{vs-g1}\}$.

2) RWNN 映射模型的输出

基于 RWNN 隐含层神经元个数 I , 使用原始训练集 $\mathbf{R}_{\text{train}} = \{\mathbf{X}_{\text{train}}, \mathbf{y}_{\text{train}}\} \in \mathbf{R}^{N \times L}$ 构建 RWNN 映射模型, 其包含输入层、输出层和单隐含层.

首先, 随机设置输入层与隐含层间神经元的连接权重 $\boldsymbol{\omega} = \{\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_I\}$ 和偏置 $\mathbf{b} = \{b_1, b_2, \dots, b_N\}^T$; 然后, 结合训练集计算隐含层的输出矩阵 \mathbf{H}^{vs-g} 和隐含层与输出层神经元的连接权重 $\boldsymbol{\beta}$ ^[35].

接着, 由 RWNN 映射函数计算虚拟样本输入对应的虚拟样本输出 \mathbf{y}_{vs-g2} 为:

$$\mathbf{y}_{vs-g2} = \Gamma_{\text{map}}(\boldsymbol{\omega}, \mathbf{b}, \mathbf{X}_{vs-g}) \boldsymbol{\beta} = \mathbf{H}^{vs-g} \boldsymbol{\beta} \quad (14)$$

最后, 获得虚拟样本集, 记为 $\mathbf{R}_{vs-g2} = \{\mathbf{X}_{vs-g}, \mathbf{y}_{vs-g2}\}$.

3.2.1.3 获得候选虚拟样本

将第 3.2.1.2 节的虚拟样本集进行混合, 获得 $\mathbf{R}_{vs-g} = \{\mathbf{R}_{vs-g1}; \mathbf{R}_{vs-g2}\} = \{\mathbf{X}_{vs-g}, \mathbf{y}_{vs-g}\}$.

本文虽然是通过在扩展域内插值生成的虚拟样本输入, 但虚拟样本输出却是通过映射模型生成的, 因此必然存在位于扩展域外的虚拟样本. 这需要根据虚拟样本输出扩展域的下限 $y_{vs-g-\text{min}}$ 和上限 $y_{vs-g-\text{max}}$ 对虚拟样本集 \mathbf{R}_{vs-g} 进行删减:

$$\mathbf{R}_{vs-d} = \{\mathbf{r}_{vs-g}^n | y_{vs-g}^n \in [y_{vs-g-\text{min}}, y_{vs-g-\text{max}}]\} \quad (15)$$

进而, 获得候选虚拟样本集 \mathbf{R}_{vs-d} , 其中 $\mathbf{r}_{vs-g}^n = \{\mathbf{x}_{vs-g}^n, \mathbf{y}_{vs-g}^n\} \in \mathbf{R}_{vs-g}$. 此外, \mathbf{R}_{vs-g} 中的样本数量 N_{vs-g} 大于等于 \mathbf{R}_{vs-d} 中的样本数量 N_{vs-d} .

3.2.2 候选虚拟样本选择

在 PSO 算法初始化过程中, 粒子的样本选择决策变量 $\mathbf{z}_{vss}^n = \{z_{vss}^1, z_{vss}^2, \dots, z_{vss}^n, \dots, z_{vss}^N\}$ 的维度设置为最大值, 以便与虚拟样本 \mathbf{R}_{vs-g} 相对应:

$$\begin{cases} \mathbf{R}_{vs-g} = \{\mathbf{r}_{vs-g}^1, \mathbf{r}_{vs-g}^2, \dots, \mathbf{r}_{vs-g}^n, \dots, \mathbf{r}_{vs-g}^{N_{vs-g}}\} \\ \mathbf{z}_{vss} = \{z_{vss}^1, z_{vss}^2, \dots, z_{vss}^n, \dots, z_{vss}^{N_{vs-g}}\} \end{cases} \quad (16)$$

式中, $z_{vss}^n \in [0, 1]$.

对 \mathbf{z}_{vss} 进行解码后, 可获得粒子所选择的虚拟样本 \mathbf{R}_{vs-s} :

$$\mathbf{R}_{vs-s} = \{\mathbf{r}_{vs-g}^n \in \mathbf{R}_{vs-g} | z_{vss}^n \geq \theta_{\text{select}}, n = 1, 2, \dots, N_{vs-g}\} \quad (17)$$

式中, θ_{select} 为虚拟样本的选择阈值, 一般设置为 0.5.

由于对 \mathbf{z}_{vss} 直接解码所获取的 \mathbf{R}_{vs-s} 中可能包含扩展域外的虚拟样本, 故需要先将 \mathbf{z}_{vss} 进行变维

度处理:

$$\mathbf{z}_{\text{vss}}^n = \begin{cases} 0, & \mathbf{r}_n \in \mathbf{R}_{\text{vs-g}} \text{ 且 } \mathbf{r}_n \notin \mathbf{R}_{\text{vs-d}} \\ z_{\text{vss}}^n, & \text{其他} \end{cases} \quad (18)$$

式 (18) 所表征的原理为: 首先, 将扩展域外的虚拟样本所对应的决策变量设置为无效; 然后, 对变维处理后的 $\mathbf{z}_{\text{vss}}^n$ 进行解码, 即从候选虚拟样本中获得虚拟样本子集 $\mathbf{R}_{\text{vs-s}}$.

3.2.3 虚拟样本评价指标计算

计算获得虚拟样本子集 $\mathbf{R}_{\text{vs-s}}$ 的评价指标, 并将其作为粒子的适应度:

$$f_{\text{num}}(\mathbf{z}) = N_{\text{vs-s}} \quad (19)$$

$$f_{\text{mod}}(\mathbf{z}) : \mathbf{R}_{\text{vs-s}} \xrightarrow{\mathbf{R}'_{\text{mix}}} \mathbf{R}'_{\text{mix}} \rightarrow f'_{\text{RF}}(\mathbf{R}_{\text{valid}}) \rightarrow F \quad (20)$$

式中, $N_{\text{vs-s}}$ 为虚拟样本集 $\mathbf{R}_{\text{vs-s}}$ 的数量, 即将虚拟样本数量作为适应度 $f_{\text{num}}(\mathbf{z})$. F 为虚拟样本集 $\mathbf{R}_{\text{vs-s}}$ 的泛化性能指标, 其计算过程为: 首先, 将原始训练集 $\mathbf{R}_{\text{train}}$ 与 $\mathbf{R}_{\text{vs-s}}$ 混合, 获得临时混合样本集 $\mathbf{R}'_{\text{mix}} = \{\mathbf{R}_{\text{train}}; \mathbf{R}_{\text{vs-s}}\}$; 然后, 基于 \mathbf{R}'_{mix} 构建临时 RF 模型 $f'_{\text{RF}}(\cdot)$ 计算验证集 $\mathbf{R}_{\text{valid}}$ 在 $f'_{\text{RF}}(\cdot)$ 上的泛化性能 F , 并将其作为适应度 $f_{\text{mod}}(\mathbf{z})$.

3.3 面向 VSG 的多目标混合优化

采用 CLPSO 算法对虚拟样本生成过程进行混合优化过程如图 2 所示, 包括种群初始化, 更新粒子速度, 更新参数决策变量, 生成候选虚拟样本, 变维更新样本选择决策变量, 选择候选虚拟样本, 计算适应度, 更新粒子个体最优、档案和样例池等阶段, 达到迭代次数后计算全局最优及获取最优虚拟样本.

种群初始化时, 首先对粒子数量 P_{num} 、迭代次数 N_{iter} 、更新阈值 N_{refresh} 、参数决策变量的上/下限等相关参数进行设定; 然后, 生成由 P_{num} 个粒子构成的种群, 随机初始化粒子的位置和速度并计算粒子的适应度; 接着, 初始化粒子的个体最优与外部档案; 最后, 计算粒子的学习概率和学习样例.

初始化种群后, 进入迭代寻优阶段. 首先, 根据式 (3) 更新粒子的速度 \mathbf{v}^p ; 然后, 根据式 (4) 更新粒子的参数决策变量的位置 $\mathbf{z}_n^p(t+1)$, 根据式 (9) 中的 $\mathbf{z}_{\text{para}}^p$ 的表征结果, 以第 3.2.1 节描述方式, 生成候选虚拟样本; 其次, 根据式 (4) 更新粒子的样本选择决策变量位置 $\mathbf{z}_{\text{vss}}^p$, 以第 3.2.2 节描述方式对候选虚拟样本进行选择, 以获得虚拟样本 $\mathbf{R}_{\text{vs-s}}$; 再依据第 3.2.3 节描述方式计算 $\mathbf{R}_{\text{vs-s}}$ 的评价指标作为粒子适应度值 $F(\mathbf{z}^p)$; 接着, 基于适应度值、根据式 (5), 更

新粒子个体最优, 并将种群搜索到的非支配解存入档案中, 并更新档案 \mathbf{A} ; 最后, 计算粒子的个体最优排序 rank^p , 并根据式 (7) 更新其学习概率 \mathbf{P}_c , 进而对迭代 N_{refresh} 次后个体最优仍未更新的粒子进行学习样例更新.

但在更新粒子学习样例时, 考虑到待优化的样本选择决策变量维数较高, 需要对 CLPSO 进行改进, 以加速虚拟样本优选过程的收敛速度. 本文首先在标准 CLPSO 采用的如式 (6) 所示的更新样例池策略的基础上, 增加样本选择决策变量向档案中粒子学习的新策略如下:

$$E_n^p = \{\mathbf{a}_n^{\text{rand}} | \mathbf{a}^{\text{rand}} \in \mathbf{A}\} \quad (21)$$

然后, 依据上述步骤不断进行迭代寻优, 在达到最大迭代次数 N_{iter} 后, 依据式 (22) 计算档案 \mathbf{A} 中粒子适应度的评估指标 ρ_i :

$$\rho_i = \frac{f_{\text{mod}}(\phi) - f_{\text{mod}}(\mathbf{a}^i)}{f_{\text{num}}(\mathbf{a}^i)}, \quad \mathbf{a}^i \in \mathbf{A} \quad (22)$$

式中, ρ_i 为全局最优粒子选择指标, 表示虚拟样本的综合评价指标; $f_{\text{mod}}(\phi)$ 表示无虚拟样本情况下, 原训练集的泛化性能指标 F ; \mathbf{a}^i 表示档案 \mathbf{A} 中的非支配解.

最后, 将档案中 ρ_i 值最大的粒子作为全局最优, 对全局最优的样本选择决策变量进行变维度解码后, 获得最优虚拟样本 \mathbf{R}_{vs} .

基于多目标 PSO 混合优化的 VSG 算法伪代码见算法 2.

算法 2. 基于多目标 PSO 混合优化的 VSG 算法伪代码

- 1) 初始化算法参数及种群;
- 2) For $n = 1$ to N_{iter} ;
- 3) For $p = 1$ to P_{num} ;
- 4) 更新粒子速度 \mathbf{v}^p ;
- 5) 更新粒子的参数决策变量的位置 $\mathbf{z}_{\text{para}}^p$; //生成候选虚拟样本;
- 6) $\gamma_{\text{extend}} \leftarrow z_{\text{MTD}}$; // 粒子的参数决策变量为 MTD 扩展率赋值;
- 7) 计算输出扩展域的上/下限 $y_{\text{vsg-max}}$ 和 $y_{\text{vsg-min}}$;
- 8) 计算输入扩展域的上/下限 $\mathbf{x}_{\text{vsg-max}}$ 和 $\mathbf{x}_{\text{vsg-min}}$;
- 9) 在小样本空间 $(\mathbf{x}_{\text{min}}, \mathbf{x}_{\text{max}})$ 、扩展空间 $(\mathbf{x}_{\text{vsg-min}}, \mathbf{x}_{\text{min}})$ 和 $(\mathbf{x}_{\text{max}}, \mathbf{x}_{\text{vsg-max}})$ 中, 分别进行等间隔插值, 以获得虚拟样本输入 $\mathbf{X}_{\text{equal}}$;
- 10) 在 $(\mathbf{x}_{\text{vsg-min}}, \mathbf{x}_{\text{vsg-max}})$ 空间中, 进行随机插值, 以获得虚拟样本输入 \mathbf{X}_{rand} ;
- 11) 获得虚拟样本输入 $\mathbf{X}_{\text{vs-g}} = \{\mathbf{X}_{\text{equal}}; \mathbf{X}_{\text{rand}}\}$;
- 12) $L_F \leftarrow z_{\text{RF}}^1$, $\theta_{\text{leaf}} \leftarrow z_{\text{RF}}^2$; //粒子的决策变量为 RF 参

数赋值;

13) 用算法 1 构建 RF 映射模型, 并计算 \mathbf{X}_{vs-g} 对应的虚拟样本输出 \mathbf{y}_{vs-g1} ;

14) $I \leftarrow z_{RWNN}$; //粒子的参数决策变量为 RWNN 参数赋值;

15) 构建 RF 映射模型, 并计算 \mathbf{X}_{vs-g} 对应的虚拟样本输出 \mathbf{y}_{vs-g2} ;

16) 获得虚拟样本 $\mathbf{R}_{vs-g} = \{\mathbf{R}_{vs-g1}; \mathbf{R}_{vs-g2}\}$;

17) 对 \mathbf{R}_{vs-g} 进行删减, 获得候选虚拟样本 \mathbf{R}_{vs-d} ;

18) 更新粒子的样本选择决策变量的位置 \mathbf{z}_{vss}^p ; //候选虚拟样本选择;

19) 对粒子样本选择决策变量的位置 \mathbf{z}_{vss}^p 进行变维度处理, 获得 \mathbf{z}_{vss}^p ;

20) 对 \mathbf{z}_{vss}^p 进行解码, 以获得虚拟样本子集 \mathbf{R}_{vs-s} ; //适应度计算;

21) $f_{\text{num}}(\mathbf{z}) \leftarrow N_{vs-s}$;

22) 使用 $\mathbf{R}'_{\text{mix}} = \{\mathbf{R}_{\text{train}}; \mathbf{R}_{vs-s}\}$, 根据算法 1, 构建 RF 模型 $f'_{\text{RF}}(\cdot)$;

23) $f_{\text{mod}}(\mathbf{z}) \leftarrow F \leftarrow f'_{\text{RF}}(\mathbf{R}_{\text{valid}})$;

24) 更新个体最优 \mathbf{d}^p 并令 $n_{\text{refresh}} = 0$. 若 \mathbf{d}^p 未更新, 则 $n_{\text{refresh}}++$;

25) End for;

26) 更新档案 \mathbf{A} ;

27) 依据粒子个体最优 \mathbf{d}^p , 计算粒子排序 rank^p ;

28) If $n_{\text{refresh}} \geq N_{\text{refresh}}$; //粒子 p 的个体最优在 N_{refresh}

次迭代后仍未被更新;

29) For $n = 1$ to N_d ;

30) If $\text{rand} < P_c^p$;

31) $E_n^p = \{\mathbf{d}_n^{p'} | \min(f(\mathbf{d}^{p'})), p' = p_{\text{rand1}}, p_{\text{rand2}}\}$;

32) Else if $\text{rand} < P_c^p$ 且 $n \in [N_{\text{para}}, N_d]$;

33) $E_n^p = \{a_n^{\text{rand}} | a^{\text{rand}} \in \mathbf{A}\}$;

34) End if;

35) End for;

36) End if;

37) End for;

38) 计算档案 \mathbf{A} 中粒子的 ρ 指标, 获得全局最优;

39) 对全局最优的样本选择决策变量进行变维度解码, 获得最优虚拟样本 \mathbf{R}_{vs} .

4 仿真验证及工业应用

基于 UCI 平台的 2 个基准数据集, 设计不同的小样本集生成虚拟样本, 对本文 VSG 方法进行验证. 通过增加虚拟样本后所构建模型的泛化性能和虚拟样本的分布情况, 验证本文方法的有效性. 进一步, 基于 MSWI 过程 DXN 排放浓度数据生成虚拟样本, 构建软测量模型.

本文进行算法仿真验证的计算机软硬件配置为 Windows7 操作系统, Matlab2021, Inter Corei7 处理器, 32 GB 内存.

4.1 评价指标综述

本文定义指标 η 用于评价小样本与数据整体间的分布相似度. 定义数据集 \mathbf{S}_1 和 \mathbf{S}_2 的分布相似度如下:

$$\eta = \frac{1}{N_{\text{attr}}} \sum_{i=1}^{N_{\text{attr}}} D_H(p_{s_1}^i \| p_{s_2}^i) \quad (23)$$

式中, N_{attr} 表示数据集 \mathbf{S}_1 和 \mathbf{S}_2 的属性数量; $p_{s_1}^i$ 和 $p_{s_2}^i$ 分别表示 $\mathbf{s}_1^i \in \mathbf{S}_1$ 和 $\mathbf{s}_2^i \in \mathbf{S}_2$ 的概率分布; $D_H(p_{s_1}^i \| p_{s_2}^i)$ 表示数据集 \mathbf{S}_1 和 \mathbf{S}_2 属性的 Hellinger 距离, 后者是 F 散度的一种, 本文用于度量 2 个概率分布的相似度:

$$D_H(p_1 \| p_2) = \sqrt{\frac{\sum_{i=1}^N \left(\sqrt{p_1(x_i)} - \sqrt{p_2(x_i)} \right)^2}{2}} \quad (24)$$

采用均方根误差 (Root mean square error, RMSE) 作为模型泛化性能的评价指标:

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^N (y_i - \hat{y}_i)^2}{N-1}} \quad (25)$$

此外, 式 (22) 定义的用以选择全局最优的评价指标 ρ_i 用于评价不同 VSG 方法生成的虚拟样本集在数量和泛化性能改进方面的优劣.

本文面向不同实验对虚拟样本进行评价, 将其重新定义为:

$$\rho_j = \frac{E_0 - E_j}{N_j} \quad (26)$$

式中, 实验 j 的虚拟样本综合评价价值定义为 ρ_j ; E_0 为基于原始小样本建模的 RMSE; E_j 为实验 j 的 RMSE; N_j 为建模所用虚拟样本的数量, 表示虚拟样本对模型泛化性能改进的平均贡献. ρ_j 值越大, 表示虚拟样本集的质量越高.

4.2 基准数据验证

4.2.1 数据描述

本文采用的基准数据集分别为混凝土抗压强度数据集和超导临界温度数据集. 其中, 混凝土抗压强度数据集共有 1030 组数据, 包含 8 个输入变量 (水泥、高炉渣、粉煤灰、水、超塑化剂、粗骨料、细骨料和龄期) 和 1 个输出变量 (混凝土抗压强度);

超导临界温度数据集共有 21 263 组数据, 包含 81 个输入变量和 1 个输出变量 (超导临界温度).

为验证本文方法, 分别对以上 2 个数据集进行处理: 从数据集中随机选取 20、40 和 60 个样本作为训练集 (即原始小样本), 对应随机选取 20、40 和 60 个样本作为验证集, 等间隔选取 100 个样本作为测试集. 每个数据集均设计 3 个对比实验, 编号分别为 A1、A2、A3、B1、B2 和 B3. 基准数据集划分如表 2 所示, 表 2 中 η 表示上述各数据集与其原始数据集根据式 (23) 计算的分布相似度.

4.2.2 实验结果

基准数据基于多目标 PSO 混合优化的 VSG 参数设定如表 3 所示, 需要根据不同数据特征凭经验确定. 分别采用 A1、A2、A3、B1、B2 和 B3 实验数据集与本文方法进行仿真实验. 基于多目标混合优化获得的非支配解的 Pareto 前沿如图 4 所示.

图 4 中的横/纵坐标分别表示 2 个优化目标, 即虚拟样本数量 N_{vs-s} 和混合样本模型的 RMSE 值. 由 2 个数据集的 Pareto 前沿可知, 当原始训练样本数为 20 时, 虚拟样本对模型性能的提升效果最为明显. 另外, 虚拟样本数量的增加可提高模型性能, 但当虚拟样本数量超过某个阈值后, 模型性能不再明显提升.

各实验均生成 1 080 个虚拟样本, 混合优化后筛选出的虚拟样本最佳数量却存在差异, 其中 A1、A2 和 A3 的最佳数量分别为 80、128 和 150, B1、B2 和 B3 的最佳数量分别约为 20、69 和 70. 这一统计结果表明, 虚拟样本的最佳数量与其质量相关.

进一步, 对非支配解进行分析. 图 5 ~ 7 分别展示了非支配解获得的虚拟样本的建模性能指标、综合评价指标和分布相似度指标的对比情况.

图 5 分别展示了不同小样本构建的 RF 软测量模型在不同测试集上的 RMSE. 由图 5 可知, 本文方法生成的虚拟样本可提高 RF 软测量模型的泛化性能. 对于超导临界温度数据集, 混合样本构建的 RF 模型在验证集上的泛化性能弱于在测试集上的表现. 另外, 随着小样本数量的增多, 基于小样本所构建模型的测试性能整体提高, 但虚拟样本对模型泛化性能却有所下降.

图 6 给出了针对混合样本构建的 RF 软测量模型在验证集和测试集上对虚拟样本的综合评价结果. 由图 6 可知, 本文方法生成的虚拟样本均有较好的综合评价指标. 但随着原始小样本数量的增加, 所生成虚拟样本的综合评价指标明显变差. 由虚拟样本的综合评价指标可知, 超导临界温度数据集在验证集上的表现较差.

不同数据集生成的虚拟样本与全体数据的分布相似度情况如图 7 所示. 由图 7 可知, 本文方法生成的虚拟样本能够改善小样本与全体数据的分布相似度, 当小样本数量为 20 时, 分布相似度改善效果最为明显. 另外, 小样本数量的增加会大大提高它与全体数据的分布相似度, 但本文方法很难对分布相似度指标进行改善. 其中当小样本数量为 60 时, 虚拟样本对分布相似度几乎未得到改善, 甚至破坏了原有分布.

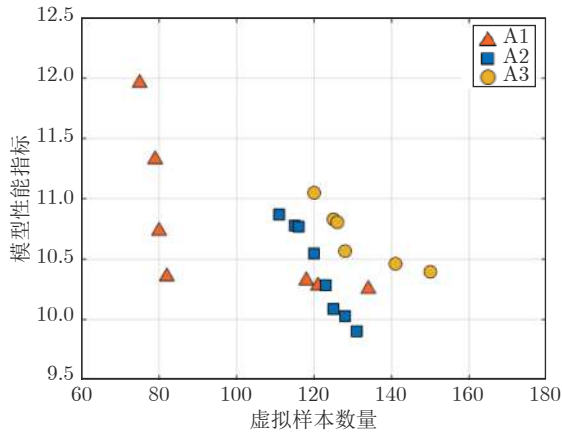
根据式 (26) 定义的综合评价指标 ρ 从档案中

表 2 基准数据集划分
Table 2 Benchmark data set partitioning

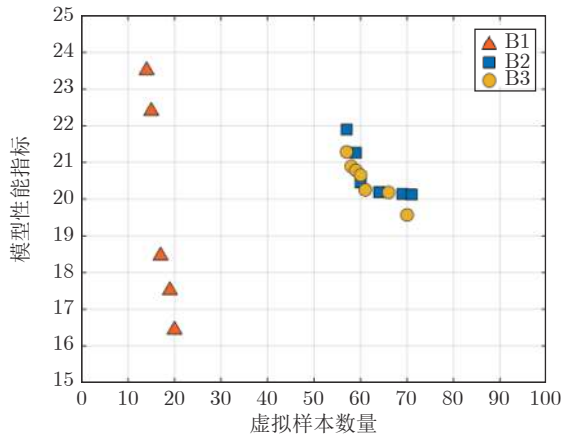
数据集	特征数	训练集		验证集		测试集		数据集编号
		数量	η	数量	η	数量	η	
混凝土抗压强度	8	20	0.3327	20	0.3598	100	0.1255	A1
		40	0.2444	40	0.2628			A2
		60	0.1853	60	0.2070			A3
超导临界温度	81	20	0.3351	20	0.3388	100	0.1538	B1
		40	0.2309	40	0.2423			B2
		60	0.1949	60	0.1966			B3

表 3 基准数据基于多目标 PSO 混合优化的 VSG 参数设定
Table 3 Parameter setting of VSG based on hybrid optimization with multi-objective PSO for benchmark data

数据集	F	N_{iter}	λ	K	z_{MTD}	z	z_{RF}^2	z_{RWNN}
混凝土抗压强度	30	30	3	30	(0, 1)	(1, 6)	(2, 10)	(3, 20)
超导临界温度	30	30	3	50	(0, 1)	(1, 30)	(2, 10)	(3, 20)



(a) Concrete compression strength dataset



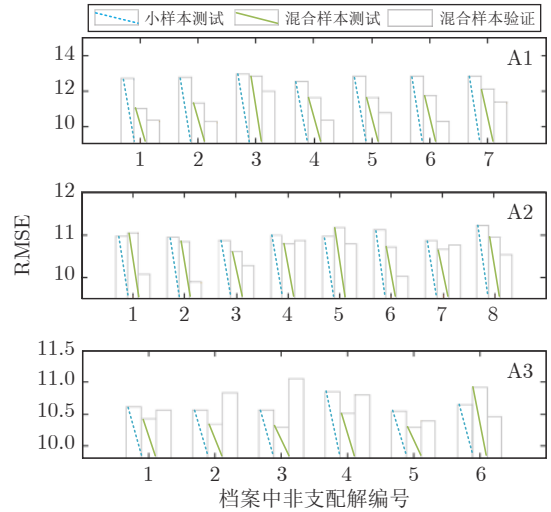
(b) Superconducting critical temperature dataset

图 4 非支配解的 Pareto 前沿

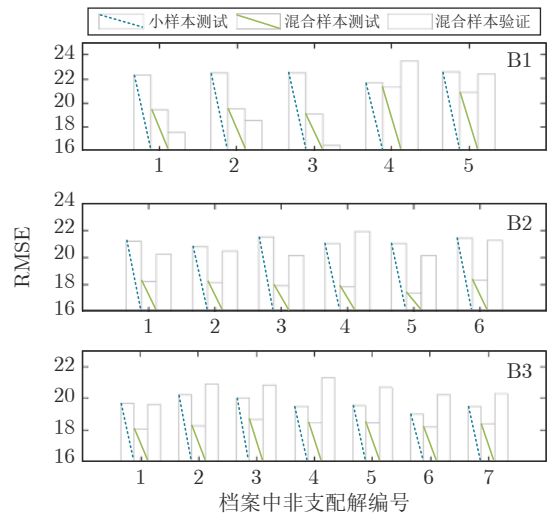
Fig. 4 Pareto front of non-dominant solutions

选取全局最优, 基准数据基于多目标 PSO 混合优化获得的最优虚拟样本如表 4 所示. 表 4 中 X_{vs} 和 y_{vs} 分别为最优虚拟样本的输入和输出. 其中, 数据集 A1 选取 5 个虚拟样本, B1 选取 5 个虚拟样本的前 8 个输入和输出. 基准数据原始样本输入/输出范围如表 5 所示.

本文方法在不同数据集上的全局最优结果包括超参数最优解、虚拟样本数量、混合样本构建的 RF 模型在验证集和测试集上的平均 RMSE、平均综合评价价值和混合样本的分布相似度指标, 基准数据基于多目标 PSO 混合优化的全局最优解的统计结果如表 6 所示. 由表 6 可知, 超参数优化结果中, 扩展率 γ_{extend} 值随数据集变化, 其受训练集和验证集分布情况的综合影响. 统计结果表明, 各数据集均进行了明显地域扩展; 小样本数量影响虚拟样本的最佳数量, 间接说明虚拟样本的最佳数量与其质量相关; 在原始训练集中加入虚拟样本构建的 RF



(a) Concrete compression strength dataset



(b) Superconducting critical temperature dataset

图 5 非支配解的建模性能指标对比

Fig. 5 Comparison of modeling performance indexes of non-dominant solutions

模型在验证集和测试集上均有较好表现, 比小样本建模的性能均有所提升. 当小样本数量为 20 时, 生成的虚拟样本最佳数量分别为 82 和 20, 建模平均测试 RMSE 分别为 11.59 和 18.05, 比小样本建模分别提升了 10.50% 和 21.73%; 虚拟样本的综合评价指标随小样本数量的增多而逐渐减小, 即虚拟样本对模型性能的提升随小样本数量的增多而变得更加困难; 同时, 混合样本与原始数据分布相似度也有较明显改善, 特别是数据集 A1 和 B1, 建模所用样本与原始数据分布相似度分别改善了 29.25% 和 38.05%.

图 8 分别给出了数据集 A 和 B 在测试集上的

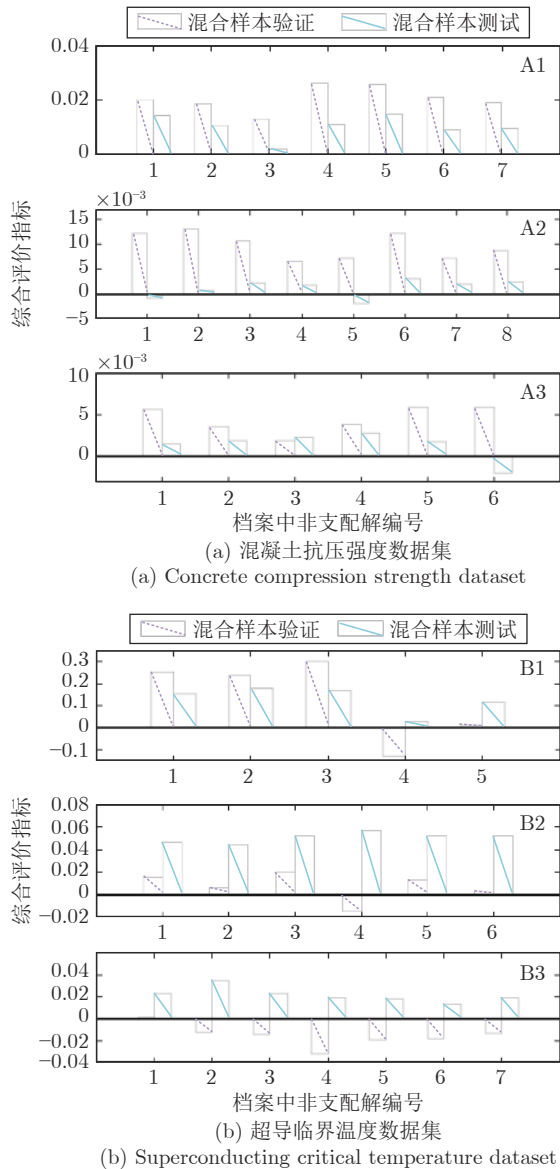


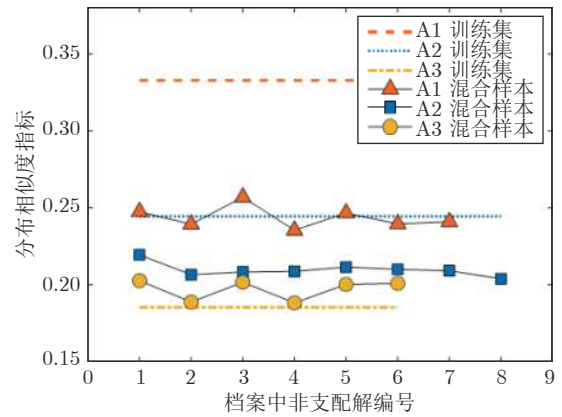
图 6 非支配解的综合评价指标对比

Fig. 6 Comparison of comprehensive evaluation indexes of non-dominant solutions

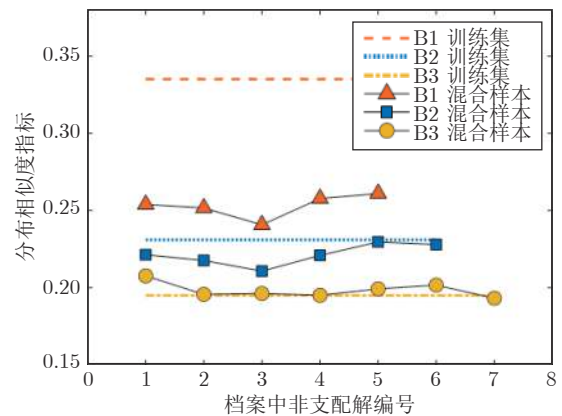
预测输出. 由图 8 可知, 混合样本建模的测试集预测输出对期望输出具有良好的拟合度, 但其精度还有提升空间.

4.2.3 方法比较

本文 MoHo-VSG 与其他 VSG 进行对比. 为了验证本文方法比其他方法更具优越性, 本节只在样本数量为 20 的 A1 和 B1 小样本集进行实验对比. 实验过程为: 首先, 采用 A1 和 B1 小样本集分别生成虚拟样本; 然后, 将其与原始小样本混合以构建模型; 最后, 所有实验重复 30 次, 并计算相应评价指标, 基准数据不同 VSG 方法的对比统计结果如表 7 所示.



(a) Concrete compression strength dataset



(b) Superconducting critical temperature dataset

图 7 非支配解的分布相似性对比

Fig. 7 Comparison of distribution similarity of non-dominant solutions

表 7 中 N-VSG^[29] 表示非线性插值的 VSG 法; M-VSG^[31] 表示线性与非线性结合的混合插值 VSG 方法; PSO-VSG^[34] 表示基于 PSO 优化生成的 VSG 方法; MP-VSG^[35] 表示基于插值并经 PSO 优化选择的 VSG 方法. 由表 7 可知, 本文方法在虚拟样本数量最少情况下, 混合样本构建的 RF 模型具有更好的泛化性能, 其在测试集上的 RMSE 和最优值最小, 表明本文方法生成的虚拟样本在提高模型泛化性能的同时, 也具有较好的稳定性. 本文方法生成的虚拟样本综合评价指标最大, 表明本文方法生成的虚拟样本具有更高质量, 即每个虚拟样本对模型性能提升的贡献更大; 本文方法生成的虚拟样本与训练集混合后的分布相似性最小, 表明其分布更符合全体数据分布.

本文所采用的 CLPSO 算法对 VSG 结果的影响主要体现在: 1) 种群的粒子数量 P_{num} 和迭代次数 N_{iter} 是对可行域进行充分搜索的基础条件, 两者的

表 4 基准数据基于多目标 PSO 混合优化获得的最优虚拟样本

Table 4 Optimal virtual samples obtained based on multi-objective PSO hybrid optimization for benchmark data

数据集	X_{vs}								y_{vs}
A1	396.50	117.40	0	176.40	11.42	876.70	796.90	60.23	58.83
	200.50	16.35	115.80	161.60	8.27	1071.70	809.90	17.23	29.23
	240.90	0	100.30	183.50	5.87	977.30	852.40	14.00	18.25
	272.40	56.58	0	199.00	0	965.00	786.90	37.38	12.62
	347.40	0	0	190.80	0	1116.40	718.20	15.08	3.42
B1	5.69	95.64	60.78	69.89	36.85	1.48	1.41	182.20	26.79
	4.08	77.39	51.82	60.19	35.09	1.22	1.27	121.40	95.32
	4.00	76.44	50.35	59.37	34.71	1.20	1.29	121.30	80.12
	4.46	82.72	56.99	64.52	36.03	1.30	1.09	131.20	51.89
	3.54	83.97	60.06	66.37	43.11	1.07	0.97	99.90	6.38

表 5 基准数据原始样本输入/输出范围

Table 5 Input/output range of original samples for benchmark data

数据集	输入									输出
A1	最小值	102.0	0	0	121.8	0	801.0	594.0	1.0	2.3
	最大值	540.0	359.4	200.1	247.0	32.2	1145.0	992.6	365.0	82.6
B1	最小值	1.0	6.9	6.4	5.3	2.0	0	0	0	0
	最大值	9.0	209.0	209.0	209.0	209.0	2.0	2.0	208.0	185.0

表 6 基准数据基于多目标 PSO 混合优化的全局最优解的统计结果

Table 6 Statistical results of global optimal solution based on hybrid optimization with multi-objective PSO for benchmark data

数据集	超参数				虚拟样本数量	验证集		测试集		混合样本 η
	γ_{extend}	L_F	θ_{leaf}	I		平均 RMSI	平均 ρ	平均 RMSI	平均 ρ	
A1	0.6033	3	9	18	82	10.36	0.026	11.59	0.012	0.2354
A2	0.6245	6	5	19	128	10.03	0.012	10.73	0.003	0.2099
A3	0.6528	6	9	20	150	10.40	0.006	10.28	0.002	0.2002
B1	0.3951	5	5	16	20	16.44	0.300	19.07	0.169	0.2407
B2	0.4892	8	6	14	69	20.14	0.019	17.86	0.051	0.2118
B3	0.6775	19	6	15	70	19.57	0	18.05	0.023	0.2076

乘积代表了粒子到达可行域的位置数. 当 P_{num} 和 N_{iter} 值过小时, 种群未收敛至全局最优; 当 P_{num} 和 N_{iter} 值过大时, 种群收敛至全局最优后继续迭代会浪费较多算力. 所以 P_{num} 和 N_{iter} 值需结合全局收敛性能和 VSG 数据进行确定. 2) 学习样例引导着粒子的搜索方向和步长, 样例池的更新阈值 N_{refresh} 决定了学习样例的更新频率, 间接决定算法的搜索能力和收敛性. 若粒子全局最优经 N_{refresh} 次未变, 需要通过更新学习样例而引导粒子跳出个体最优. 当 N_{refresh} 过大时, 粒子长期向旧学习样例进行学习会导致种群全局搜索能力下降; 当 N_{refresh} 过小时, 粒子不断向新学习样例进行学习会导致种群收敛性

变差. 所以 N_{refresh} 值的确定需考虑 VSG 数据的特性并结合迭代次数 N_{iter} 值. 另外, 变量 z_{MTD} 、 z_{RF}^1 、 z_{RF}^2 、 z_{RWNN} 的上/下限决定着种群的可行域, 影响着种群的搜索效率和结果. 当可行域过大, 则搜索效率会下降; 当可行域过小, 则可能会错失全局最优解. 所以, 它们的取值也需要根据 VSG 数据特征, 凭经验确定.

4.3 工业数据验证

4.3.1 二噁英排放过程描述

国内 MSWI 过程工艺流程图如图 9 所示. 图 9 中, 由 MSWI 过程所产生的 DXN 分别包含在灰渣、

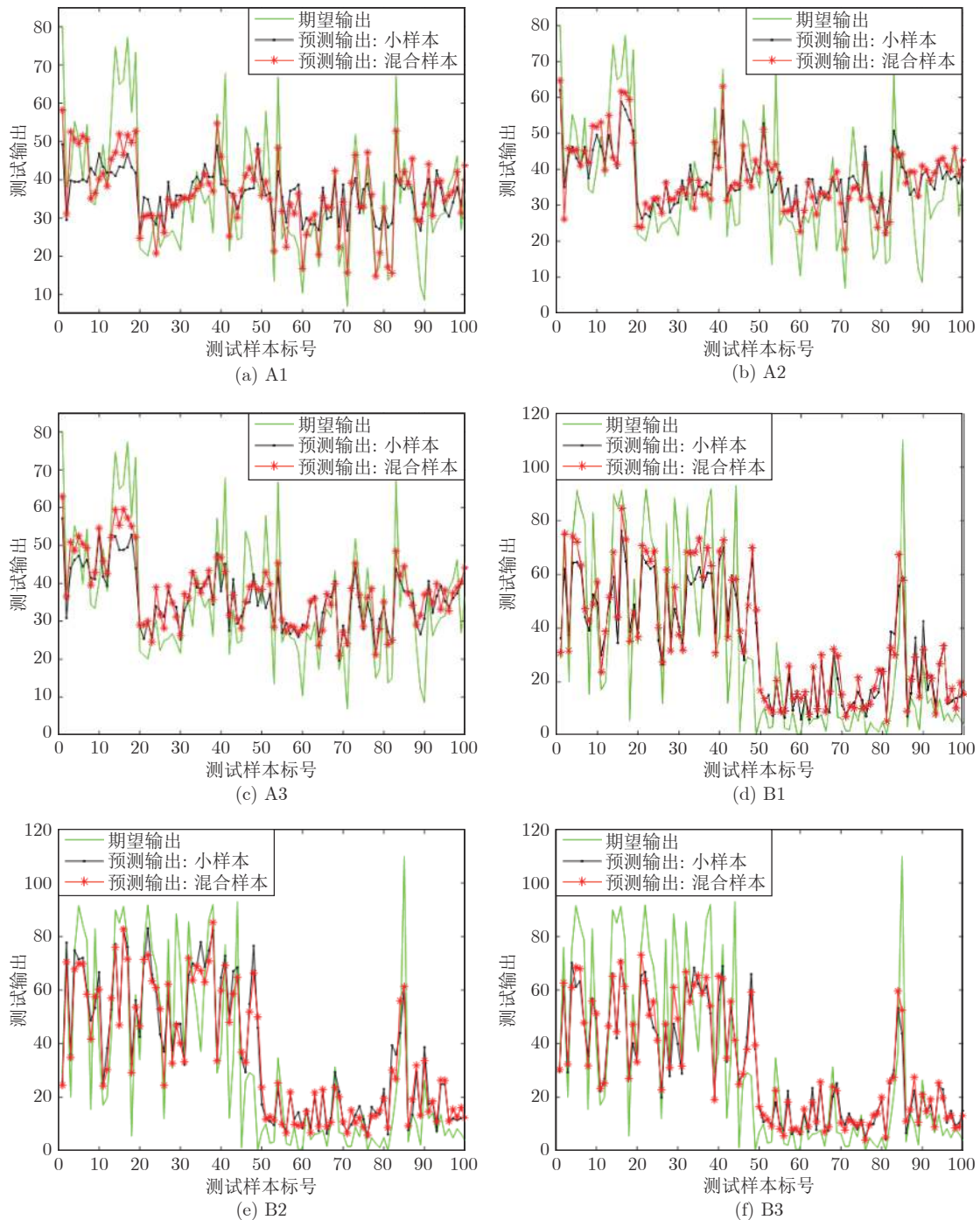


图 8 基准数据预测输出对比

Fig. 8 Comparison of prediction output for benchmark data

飞灰和烟气 3 种产物中, 其中烟气中含有的 DXN 按照工艺阶段可分为 DXN 产生时的烟气 G1、DXN 被吸附后的烟气 G2 和排放至大气的烟气 G3 三种。在机理上, DXN 的产生来源包括固废不完全燃烧和新规合成反应生成 2 类^[54]。通常, 为保证 DXN 等有毒有机物的有效分解, 在固废焚烧阶段的烟气温度应达到至少 850 °C 并保持 2 s。另外, 为减低排

放烟气中的 DXN 浓度, 在烟气处理阶段需要向反应器内喷射消石灰和活性炭, 以吸附 DXN 以及某些重金属。此外, 余热锅炉和烟气处理阶段的积灰所造成的至今机理仍不清晰的 DXN 记忆效应也会导致 DXN 排放浓度增加。上述不同阶段的过程变量均以秒为周期、由现场控制系统采集。但焚烧企业或环保部门通常以月、季或更长时间为不确定周

表 7 基准数据不同 VSG 方法的对比统计结果

Table 7 Comparative statistical results of different VSG methods for benchmark data

数据集	方法	虚拟样本数量	混合样本 η	测试 κ			测试 ρ		
				均值	方差	最优	均值 ($\times 10^1$)	方差 ($\times 10^1$)	最优 ($\times 10^{-1}$)
A1	N-VSG	219	0.2770	16.47	8.785	14.11	4.09	15.44	4.62
	M-VSG	238	0.3018	17.08	8.575	13.65	2.26	19.73	4.55
	PSO-VSG	55	0.4235	16.35	3.822	12.75	3.76	30.20	5.88
	MP-VSG	165	0.2641	14.03	4.525	12.93	6.04	9.93	7.19
	MoHo-VSG	82	0.2354	11.59	0.107	9.67	12.46	1.34	14.72
B1	N-VSG	176	0.2945	24.38	10.541	21.96	13.87	17.96	14.25
	M-VSG	281	0.3100	25.33	12.786	20.12	12.63	56.11	14.12
	PSO-VSG	36	0.3317	26.11	17.710	20.38	1.69	71.20	8.23
	MP-VSG	134	0.2513	20.84	3.452	19.47	17.43	4.37	18.89
	MoHo-VSG	20	0.2076	18.05	0.062	17.84	169.26	1.57	178.69

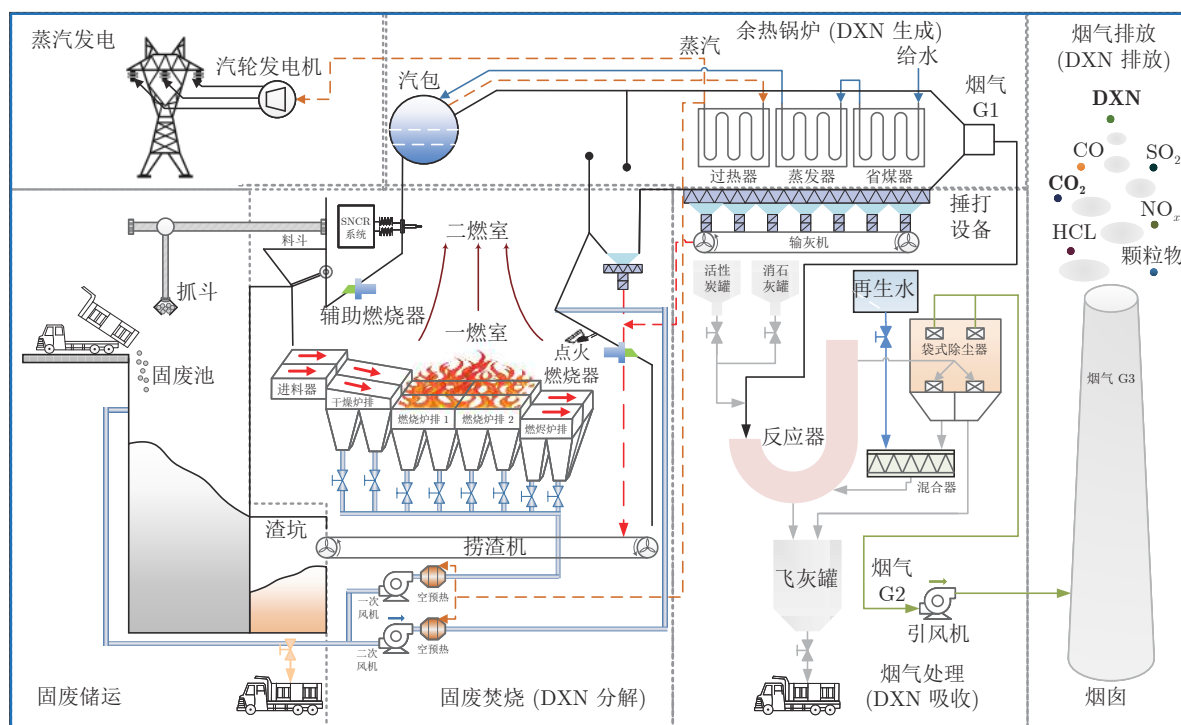


图 9 MSWI 过程工艺流程图

Fig.9 Flow chart of MSWI process

期, 离线化验烟气 G3 中 DXN 浓度, 该方法需要专门的实验室分析设备, 检测成本高且耗时长^[4]. 此外, 烟气 G3 中的易检测气体 (如 CO、HCL、SO₂ 和 NO_x 等) 浓度能够通过烟气排放连续监测系统实时检测, 并且与 DXN 浓度存在相关性. 基于指标/关联的在线间接检测方法要求先检测指示物/相关物的浓度, 再基于映射模型间接计算 DXN 排放浓度, 需要昂贵且复杂的在线分析设备, 并且存在以小时为单位的时间滞后^[1]. 因此, 有必要构

建 DXN 排放浓度软测量模型, 以实现在线实时检测.

综上所述, 烟气 G3 中的 DXN 浓度与 MSWI 过程不同阶段的过程变量相关, 并且构建 DXN 预测模型的数据 (真输入-真输出) 具有样本数量稀缺与分布不均衡、输入特征维数高等特性. Bunsan 等^[5] 结合机理和经验, 利用台湾某焚烧厂 4 年多的实际过程数据, 结合相关分析、主成分分析和人工神经网络, 从 23 个易检测变量中选取 13 个变量建立 DXN

软测量模型. Xiao 等^[55]采用炉温、锅炉出口烟气温、烟气流、SO₂、HCL 和颗粒物浓度等输入变量, 建立基于支持向量机的 DXN 排放浓度软测量模型. 针对实际 MSWI 过程变量具有数百维且不同程度地与 DXN 产生、吸收和排放有关, 乔俊飞等^[56]提出多层特征选择方法. 但是, 以上方法均是通过降低建模样本维度的方式构建软测量模型, 并未从本质上解决建模样本稀少问题, 并且未被选择的特征可能会造成信息损失. 因此, 本文采用 MoHo-VSG 用于解决 DXN 排放浓度建模问题.

4.3.2 数据描述

本文采用的工业数据源于北京某基于炉排炉的 MSWI 电厂, 涵盖了 2012 ~ 2018 年所记录的有效 DXN 排放浓度检测样本共 34 个. 将原始数据经过预处理后, 获得包含 119 维输入和 1 维输出的建模样本. 由于原始样本数量较少, 将数据集划分为训练集和验证集, 验证集同时也作为测试集, 将该数据集记为 C.

4.3.3 实验结果

DXN 数据基于多目标 PSO 混合优化的 VSG 算法参数设定如表 8 所示, 包括决策变量 z_{MTD} 、 z_{RF}^1 、 z_{RF}^2 和 z_{RWNN} 的最大/最小值.

在数据集 C 上, 对本文方法进行仿真实验, 获得非支配的 Pareto 前沿 —— DXN 排放浓度如

图 10 所示. 由图 10 可以看出, 当虚拟样本数量为 40 时, 模型泛化性能较好, 其中候选虚拟样本数量均为 918.

对非支配解进行分析. 非支配解的建模性能和综合评价指标对比如图 11 所示. 由图 11 可以看出, 本文方法生成的虚拟样本在总体上可提高模型的泛化性能; 而非支配解 4 和 5 的综合评价指标为负, 表明加入虚拟样本后, 建模性能没有得到提升, 反而降低了. 由于测试集和验证集相同, 所构建的 RF 模型表现相近, 但也存在一定差别. DXN 数据基于多目标 PSO 混合优化获得的最优虚拟样本如表 9 所示, 表 9 展示了 5 个虚拟样本的前 7 个输入和 1 个输出.

DXN 数据面向 VSG 的多目标 PSO 混合优化全局最优解如表 10 所示. 由表 10 可以看出, 超参数 γ_{extend} 较小, 表明样本域扩展程度较小, 反映了训练集与测试集的分布域较为相似. 17 个训练样本生成的虚拟样本最佳数量为 40, 混合样本构建的 RF 模型在验证集和测试集上的表现相近, 其在测试集上的平均 RMSE 为 0.023 1, 比小样本建模提升了 2.51%; 虚拟样本的综合评价指标大于 0, 但值较小, 表明所生成的虚拟样本有效用但仍需改进.

4.3.4 方法比较

本文 MoHo-VSG 与其他 VSG 进行对比. 采用

表 8 DXN 数据基于多目标 PSO 混合优化的 VSG 算法参数设定

Table 8 Parameter setting of VSG algorithm based on multi-objective PSO hybrid optimization for DXN data

参数	P_n	N_{iter}	N_{rel}	K	z_{MTD}	z_{RF}^1	z_{RF}^2	z
数据	30	30	3	50	(0, 1)	(1, 35)	(2, 10)	(3, 20)

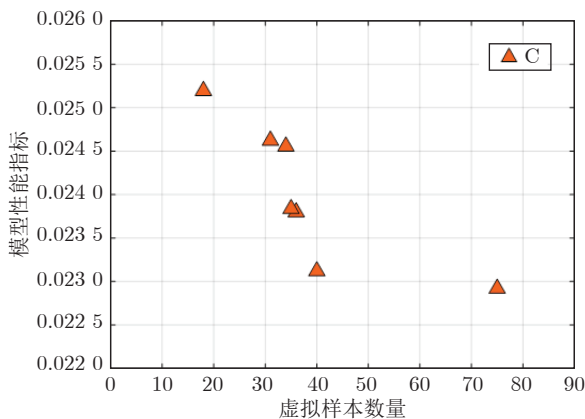


图 10 非支配解的 Pareto 前沿 —— DXN 排放浓度
Fig.10 Pareto front of non-dominated solutions —— DXN emission concentration

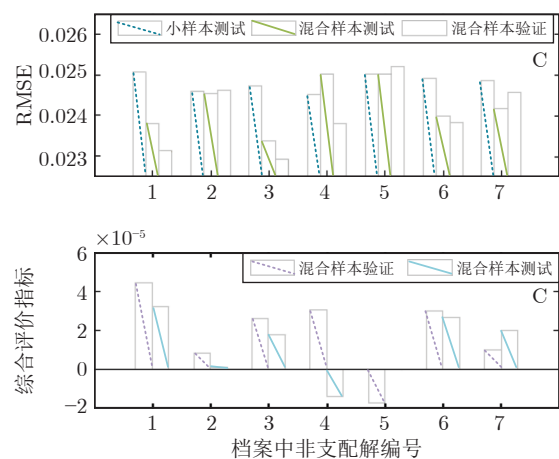


图 11 非支配解的建模性能和综合评价指标对比
Fig.11 Comparison of modeling performance indexes and comprehensive evaluation indexes of non-dominant solutions

表 9 DXN 数据基于多目标 PSO 混合优化获得的最优虚拟样本

Table 9 Optimal virtual samples obtained based on multi-objective PSO hybrid optimization for DXN data

			\mathbf{x}_{vs}	y_{vs}			
4.366	1.54	68.78	27.31	241.4	3.96	334.7	0.0289
4.206	0	68.94	28.15	222.5	3.77	306.8	0.0458
4.449	7.69	72.48	30.23	222.8	3.98	315.8	0.0685
4.432	10.00	71.83	30.00	225.9	3.99	319.5	0.0163
4.461	17.69	74.65	30.77	228.5	3.99	321.8	0.0029

表 10 DXN 数据面向 VSG 的多目标 PSO 混合优化全局最优解

Table 10 DXN data for VSG-oriented multi-objective PSO hybrid optimization global optimal solution

性能指标	最优解
超参数 γ_{extc}	0.1206
超参数 L	2
超参数 θ_{leaf}	5
超参数 l	15
虚拟样本数量	40
验证集的平均 RMSE	0.0231
验证集的平均 ρ	4.41×10^{-5}
测试集的平均 RMSE	0.0238
测试集的平均 ρ	3.18×10^{-5}
验证集, 小样本建模的 RMSE	0.0259
测试集, 小样本建模的 RMSE	0.0251

数据集 C 生成虚拟样本, 将其与原始小样本混合构建模型, 实验均重复 30 次, DXN 数据的不同 VSG 方法对比统计结果如表 11 所示. 由表 11 可以看出, 本文方法在虚拟样本数量最少情况下, 混合样本构建的 RF 模型具有更好的泛化性能, 其在测试集上的 RMSE 均值和方差较小, 表明本文方法在提升模型预测性能和稳定性上具有优势. 但是, 在 30 次重复实验中, 本文方法最优 RMSE 值不如 MP-VSG 方法. 另外, 本文方法生成的虚拟样本有较好的综合评价指标 ρ .

综上所述, 本文 MoHo-VSG 能够对 VSG 过程

的超参数和虚拟样本的选择进行混合优化, 确保优选并生成更为合理的虚拟样本, 能够有效地提高虚拟样本的质量和确定其最佳数量. 针对不同的数据集, 本文方法能进行自适应的域扩展, 并基于生成的虚拟样本优化确定其最佳数量. 生成的虚拟样本可明显提升模型泛化性能, 且具有较好的综合评价值, 也能够提高小样本与全体数据的分布相似度 η 值, 比其他 VSG 方法具有优势. 在 2 个基准数据集上进行仿真实验, 当小样本数量为 20 时, 生成的虚拟样本最佳数量分别为 82 和 20, 建模平均 RMSE 分别为 11.59 和 18.05, 比小样本建模分别提升了 10.50% 和 21.73%, 建模所用样本与原始数据的分布相似度分别改善了 29.25% 和 38.05%. 将本文方法应用于 DXN 排放浓度建模上, 17 个训练样本生成的虚拟样本最佳数量为 40, 模型在测试集上的平均 RMSE 为 0.0231, 比小样本建模提升了 2.51%.

5 结束语

针对工业过程回归建模时样本数量有限问题, 本文提出基于多目标 PSO 混合优化的 VSG 方法, 其创新性表现有以下 3 点: 1) 首次采用混合优化策略对 VSG 过程的超参数和样本选择过程进行同时优化, 确保虚拟样本的合理性和有效性; 2) 改进 CL-PSO 算法对 VSG 过程进行多目标优化, 在确保模型泛化性能的同时, 尽可能地降低虚拟样本数量, 这样既保证了虚拟样本的整体质量, 也确定了虚拟样本的最佳数量; 3) 提出新的面向虚拟样本质量的

表 11 DXN 数据的不同 VSG 方法对比统计结果

Table 11 Comparative statistical results of different VSG methods based on DXN dataset

方法	虚拟样本数量	测试集 RMS			测试集 ρ		
		均值	方差 ($\times 10^{-7}$)	最优	均值 ($\times 10^{-5}$)	方差	最优 ($\times 10^{-5}$)
N-VSG	129	0.0406	0.695	0.0262	0.19	1.94×10^{-7}	0.36
M-VSG	116	0.0403	1.331	0.0231	0.26	8.83×10^{-7}	0.53
PSO-VSG	27	0.0328	0.519	0.0245	0.56	8.44×10^{-7}	1.02
MP-VSG	68	0.0377	1.208	0.0218	1.04	5.16×10^{-7}	1.78
MoHo-VSG	40	0.0231	0.691	0.0220	3.18	4.47×10^{-7}	3.45

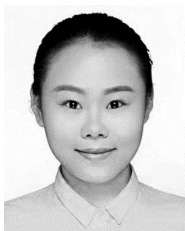
综合评价指标和分布相似度指标, 用于度量虚拟样本对建模性能的贡献度, 以及虚拟样本改善小样本分布的效果. 通过基准数据和工业数据仿真实验, 验证了本文方法的有效性.

目前, 面向工业过程小样本数据回归建模的 VSG 方法仍处于不断探索的阶段, 在如何确定样本的期望分布、如何针对不同研究领域小样本数据的特性从理论上确定虚拟样本最佳数量、如何提出更好的虚拟样本评价指标以度量虚拟样本和实际数据的差异等方向, 仍有待深入研究.

References

- Qiao Jun-Fei, Guo Zi-Hao, Tang Jian. A review on the determination of dioxin emission concentration in municipal solid waste incineration process. *Acta Automatica Sinica*, 2020, **46**(6): 1063–1089
(乔俊飞, 郭子豪, 汤健. 面向城市固废焚烧过程的二噁英排放浓度检测方法综述. *自动化学报*, 2020, **46**(6): 1063–1089)
- Chai Tian-You. Industrial process control systems: Research status and development direction. *Scientia Sinica Informationis*, 2016, **46**(8): 1003–1015
(柴天佑. 工业过程控制系统研究现状与发展方向. *中国科学: 信息科学*, 2016, **46**(8): 1003–1015)
- Arafat H A, Jijakli K, Ahsan A. Environmental performance and energy recovery potential of five processes for municipal solid waste treatment. *Journal of Cleaner Production*, 2015, **105**: 233–240
- Zhou H, Meng A, Long Y Q, Li Q H, Zhang Y G. A review of dioxin-related substances during municipal solid waste incineration. *Waste Management*, 2015, **36**(8): 106–118
- Jones P H, Degerlache J, Marti E, Mischer G, Niessen H J. The global exposure of man to dioxins: A perspective on industrial-waste incineration. *Chemosphere*, 1993, **26**: 1491–1497
- Tang Jian, Qiao Jun-Fei. Soft sensor of dioxin emission concentration in solid waste incineration process based on selective ensemble kernel learning algorithm. *Journal of Chemical Engineering and Technology*, 2019, **70**(2): 696–706
(汤健, 乔俊飞. 基于选择性集成核学习算法的固废焚烧过程二噁英排放浓度软测量. *化工学报*, 2019, **70**(2): 696–706)
- He A, Li T, Li N, Wang K, Fu H. CABNet: Category attention block for imbalanced diabetic retinopathy grading. *IEEE Transactions on Medical Imaging*, 2021, **40**(1): 143–153
- Wang Q, Wang K, Li Q, Yang Z, Jin G, Wang H. MBNN: A multi-branch neural network capable of utilizing industrial sample unbalance for fast inference. *IEEE Sensors Journal*, 2021, **21**(2): 1809–1819
- Tang Jian, Qiao Jun-Fei, Chai Tian-You, Liu Zhuo, Wu Zhi-Wei. Multi-component mechanical signal modeling based on virtual sample generation technology. *Acta Automatica Sinica*, 2018, **44**(9): 1569–1589
(汤健, 乔俊飞, 柴天佑, 刘卓, 吴志伟. 基于虚拟样本生成技术的多组分机械信号建模. *自动化学报*, 2018, **44**(9): 1569–1589)
- Lin Y S, Li D C. The generalized-trend-diffusion modeling algorithm for small data sets in the early stages of manufacturing systems. *European Journal of Operational Research*, 2010, **207**(1): 121–130
- Zhu Q X, Chen Z, Zhang X H, Rajabifard A, Chen Y. Dealing with small sample size problems in process industry using virtual sample generation: A Kriging-based approach. *Soft Computing*, 2020, **24**(9): 6889–6902
- Zhang T, Chen J, Xie J, Pan T. SASLN: Signals augmented self-taught learning networks for mechanical fault diagnosis under small sample condition. *IEEE Transactions on Instrumentation and Measurement*, 2021, **70**: 1–11
- Poggio T, Vetter T. Recognition and structure from one 2D model view: Observations on-prototypes, object classes and symmetries. *Laboratory Massachusetts Institute of Technology*, 1992: Article No. 1347
- Li D C, Lin L S, Chen C C, Yu W H. Using virtual samples to improve learning performance for small datasets with multimodal distributions. *Soft Computing*, 2019, **23**(22): 11883–11900
- Niyogi P, Girosi F, Poggio T. Incorporating prior information in machine learning by creating virtual examples. *Proceedings of the IEEE*, 1998, **86**(11): 2196–2209
- Li D C, Hsu H C, Tsai T I, Te J L, Susan C H. A new method to help diagnose cancers for small sample size. *Expert Systems With Applications*, 2007, **33**(2): 420–424
- Zhu Y, Yao J. A novel reliability assessment method based on virtual sample generation and failure physical model. In: Proceedings of the 12th International Conference on Reliability, Maintainability, and Safety. Shanghai, China: 2018. 99–102
- Schlkopf B, Simard P, Smola A J, Vapnik V. Prior knowledge in support vector kernels. In: Proceedings of Neural Information Processing Systems. Denver, USA: 1997. 640–646
- Cai W D, Ma B, Zhang L, Han Y M. A pointer meter recognition method based on virtual sample generation technology. *Measurement*, 2020, **163**: Article No. 107962
- Gang H, Yuan X, Wei Z, Shi Y. An effective method for face recognition by creating virtual training samples based on pixel processing. In: Proceedings of the 10th International Conference on Intelligent Human-Machine Systems and Cybernetics. Hangzhou, China: 2018. 177–180
- Luo J, Tjahjadi T. Multi-set canonical correlation analysis for 3D abnormal gait behaviour recognition based on virtual sample generation. *IEEE Access*, 2020, **8**: 32485–32501
- Li D C, Lin Y S. Using virtual sample generation to build up management knowledge in the early manufacturing stages. *European Journal of Operational Research*, 2006, **175**(1): 413–434
- Li D C, Lin L S. A new approach to assess product lifetime performance for small data sets. *European Journal of Operational Research*, 2013, **230**(2): 290–298
- Lin L S, Li D C, Yu W H, Hsueh Y M. Generating multi-modality virtual samples with soft DBSCAN for small dataset learning. In: Proceedings of the 3rd International Conference on Applied Computing and Information Technology/2nd International Conference on Computational Science and Intelligence. Okayama, Japan: 2015. 363–368
- Zhang X H, Xu Y, He Y L, Zhu Q X. Novel manifold learning based virtual sample generation for optimizing soft sensor with small data. *ISA Transactions*, 2021, **109**: 229–241
- Chen Z S, Zhu Q X, Xu Y, He Y L, Nagy Z K. Integrating virtual sample generation with input-training neural network for solving small sample size problems: Application to purified terephthalic acid solvent system. *Soft Computing*, 2021, **25**(8): 6489–6504
- Li D C, Chen C C, Chang C J, Lin W K. A tree-based-trend-diffusion prediction procedure for small sample sets in the early stages of manufacturing systems. *Expert Systems With Applications*, 2012, **39**(1): 1575–1581

- 28 Zhu B, Chen Z S, Yu L A. A novel small sample mage-trend-diffusion technology. *Journal of Chemical Industry and Technology*, 2016, **67**(3): 820–826
- 29 He Y L, Wang P J, Zhang M Q, Zhu Q X, Xu Y A. A novel and effective nonlinear interpolation virtual sample generation method for enhancing energy prediction and analysis on small data problem: A case study of ethylene industry. *Energy*, 2018, **147**: 418–427
- 30 Zhu Bao, Qiao Jun-Fei. Virtual sample generation method based on AANN feature scaling and its process modeling application. *Computer and Applied Chemistry*, 2019, **36**(4): 304–307 (朱宝, 乔俊飞. 基于 AANN 特征缩放的虚拟样本生成方法及其过程建模应用. *计算机与应用化学*, 2019, **36**(4): 304–307)
- 31 Qiao J F, Guo Z H, Tang J. Virtual sample generation method based on improved megatrend diffusion and hidden layer interpolation and its application. *Journal of Chemical Industry and Engineering*, 2020, **71**(12): 5681–5695
- 32 Tang J, Jia M, Liu Z, Chai T Y, Yu W. Modeling high dimensional frequency spectral data based on virtual sample generation technique. In: Proceedings of the International Conference on Information and Automation. Lijiang, China: 2015. 1090–1095
- 33 Li D C, Wen I. A genetic algorithm-based virtual sample generation technique to improve small data set learning. *Neurocomputing*, 2014, **143**: 222–230
- 34 Chen Z S, Zhu B, He Y L, Yu L A. A PSO based virtual sample generation method for small sample sets: Applications to regression datasets. *Engineering Applications of Artificial Intelligence*, 2016, **59**: 236–243
- 35 Tang Jian, Wang Dan-Dan, Guo Zi-Hao, Qiao Jun-Fei. Prediction of dioxin emission concentration in urban solid waste incineration process based on virtual sample optimization selection. *Journal of Beijing University of Technology*, 2021, **47**(5): 431–443 (汤健, 王丹丹, 郭子豪, 乔俊飞. 基于虚拟样本优化选择的城市固废焚烧过程二噁英排放浓度预测. *北京工业大学学报*, 2021, **47**(5): 431–443)
- 36 Tang Jian, Xia Heng, Qiao Jun-Fei, Guo Zi-Hao. Research on deeply integrated forest regression modeling method and its application. *Journal of Beijing University of Technology*, 2021, **47**(11): 1219–1229 (汤健, 夏恒, 乔俊飞, 郭子豪. 深度集成森林回归建模方法及应用研究. *北京工业大学学报*, 2021, **47**(11): 1219–1229)
- 37 Liang J J, Qin A K, Suganthan P N, Baskar S. Comprehensive learning particle swarm optimizer for global optimization of multi-modal functions. *IEEE Transactions on Evolutionary Computation*, 2006, **10**(3): 281–295
- 38 Tang J, Zhang J, Yu G, Zhang W P, Yu W. Multi-source latent feature selective ensemble modeling approach for small-sample high-dimension process data in application. *IEEE Access*, 2020, **8**: 148475–148488
- 39 Lin Yue, Liu Ting-Zhang, Wang Zhe-He. Optimization of virtual sample generating quantity with two kinds of upper limit conditions. *Journal of Guangxi Normal University (Natural Science Edition)*, 2019, **37**(1): 142–148 (林越, 刘廷章, 王哲河. 具有两类上限条件的虚拟样本生成数量优化. *广西师范大学学报 (自然科学版)*, 2019, **37**(1): 142–148)
- 40 Vallejo M, Espriella C, Gómez-Santamaría J, Ramírez-Barrera A F, Delgado-Trejos E. Soft metrology based on machine learning: A review. *Measurement Science and Technology*, 2020, **31**(3): Article No. 32001
- 41 Tang Jian, Qiao Jun-Fei, Xu Zhe, Guo Zi-Hao. Soft measurement of dioxin emission concentration in municipal solid waste incineration process based on feature reduction and selective integration algorithm. *Control Theory & Applications*, 2021, **38**(1): 110–120 (汤健, 乔俊飞, 徐喆, 郭子豪. 基于特征约简与选择性集成算法的城市固废焚烧过程二噁英排放浓度软测量. *控制理论与应用*, 2021, **38**(1): 110–120)
- 42 Zhong K, Han M, Han B. Data-driven based fault prognosis for industrial systems: A concise overview. *IEEE/CAA Journal of Automatica Sinica*, 2020, **7**(2): 330–345
- 43 Zhu Bao. Virtual Sample Generation Technology and Modeling Application [Ph.D. dissertation], Beijing University of Chemical Technology, China, 2017. (朱宝. 虚拟样本生成技术及建模应用研究 [博士论文], 北京化工大学, 中国, 2017.)
- 44 Li D C, Lin L S, Peng L J. Improving learning accuracy by using synthetic samples for small datasets with non-linear attribute dependency. *Decision Support Systems*, 2014, **59**: 286–295
- 45 Chen Z S, Zhu B, He Y L, Yu L A. A PSO based virtual sample generation method for small sample sets: Applications to regression datasets. *Engineering Applications of Artificial Intelligence*, 2017, **59**: 236–243
- 46 Wang Y Q, Wang Z Y, Sun J Y, Zhang J J, Zissimos M. Gray bootstrap method for estimating frequency-varying random vibration signals with small samples. *Chinese Journal of Aeronautics*, 2014, **27**(2): 383–389
- 47 Hong W C, Li M W, Geng J, Zhang Y. Novel chaotic bat algorithm for forecasting complex motion of floating platforms. *Applied Mathematical Modelling*, 2019, **72**: 425–443
- 48 Bloch G, Lauer F, Colin G, Chamailard Y. Support vector regression from simulation data and few experimental samples. *Information Sciences*, 2008, **178**(20): 3813–3827
- 49 Thomas P T, Edward A P. Small sample reliability growth modeling using a grey systems model. *Grey Systems Theory and Application*, 2018, **8**(3): 246–271
- 50 Shapiai M I, Ibrahim Z, Khalid M, Jau L W, Pavlovic V, Watada J. Function and surface approximation based on enhanced kernel regression for small sample set. *International Journal of Innovative Computing, Information & Control: IJCI-CIC*, 2011, **7**(10): 5947–5960
- 51 Dai Z, Wei H, Li X, Lv M. Validation of issile simulation model based on Bayesian theory with extreme small sample. In: Proceedings of the 3rd International Conference on Electron Device and Mechanical Engineering. Suzhou, China: 2020. 683–686
- 52 Hou Y, Zheng E, Guo W, Xiao Q, Xu Z. Learning Bayesian network parameters with small data set: A parameter extension under constraints method. *IEEE Access*, 2020, **8**: 24979–24989
- 53 Yu Xu, Yang Jing, Xie Zhi-Qiang. Research on virtual sample generation technology. *Computer Science*, 2011, **38**(3): 16–19 (于旭, 杨静, 谢志强. 虚拟样本生成技术研究. *计算机科学*, 2011, **38**(3): 16–19)
- 54 Bunsan S, Chen W Y, Chen H W, Grisdanurak N. Modeling the dioxin emission of a municipal solid waste incinerator using neural networks. *Chemosphere*, 2013, **92**: 258–264
- 55 Xiao X D, Lu J W, Hai J. Prediction of dioxin emissions in flue gas from waste incineration based on support vector regression. *Renewable Energy Resources*, 2017, **35**(8): 1107–1114
- 56 Qiao Jun-Fei, Guo Zi-Hao, Tang Jian. Soft sensing of dioxin emission concentration in solid waste incineration process based on multi-layer feature selection. *Information and Control*, 2021, **50**(1): 75–87 (乔俊飞, 郭子豪, 汤健. 基于多层特征选择的固废焚烧过程二噁英排放浓度软测量. *信息与控制*, 2021, **50**(1): 75–87)



王丹丹 北京工业大学信息学部硕士研究生. 主要研究方向为基于虚拟样本生成的小样本数据建模.

E-mail: wangdandan@emails.bjut.edu.cn

(**WANG Dan-Dan** Master student at the Faculty of Information Technology, Beijing University of Technology. Her main research interest is small sample data modeling based on virtual sample generation.)



汤健 北京工业大学信息学部教授. 主要研究方向为小样本数据建模, 城市固废处理过程智能控制. 本文通信作者.

E-mail: freeflytang@bjut.edu.cn

(**TANG Jian** Professor at the Faculty of Information Technology, Beijing University of Technology. His research interest covers small sample data modeling and intelligent control of municipal solid waste treatment process. Corresponding author of this paper.)



夏恒 北京工业大学信息学部博士研究生. 主要研究方向为小样本数据建模和城市固废焚烧过程二噁英排放预测.

E-mail: xiaheng@emails.bjut.edu.cn

(**XIA Heng** Ph.D. candidate at the Faculty of Information Technology, Beijing University of Technology. His research interest covers small sample data modeling and dioxin emission prediction of the municipal solid waste incineration process.)



乔俊飞 北京工业大学信息学部教授. 主要研究方向为污水处理过程智能控制, 神经网络结构设计与优化.

E-mail: junfei@bjut.edu.cn

(**QIAO Jun-Fei** Professor at the Faculty of Information Technology, Beijing University of Technology. His research interest covers intelligent control of waste water treatment process and structure design and optimization of neural networks.)