DOI: 10.19650/j.cnki.cjsi.J2209392

基于 Bagging 半监督深度森林回归的二噁英排放浓度软测量*

徐 雯,汤 健,夏 恒,乔俊飞
(北京工业大学信息学部 北京 100124)

摘 要:城市固废焚烧(MSWI)过程产生的副产品之一是被称为“世纪之毒”的二噁英(DXN),受限于其排放浓度检测技术难度以及时间与经济成本等因素,难以获得足量的有标记样本用于构建 DXN 排放浓度软测量模型。为有效利用现场控制系统采集的大量无标记样本,同时解决传统浅层学习模型泛化性能较差的问题,提出了基于 Bagging 半监督深度森林回归(DFR)的 DXN 排放浓度软测量方法。首先,基于 Bagging 机制以重采样原始标记数据集的方式获得多个训练子集,并构建具有差异性的多个随机森林(RF)模型;接着,将 RF 模型迭代更新、近邻集合选择和性能评估策略相结合用于获得高置信度伪标记样本;最后,基于伪标记和原始标记样本集构建 DFR 模型。采用北京某 MSWI 电厂的实际 DXN 检测数据验证了所提方法的有效性,结果表明,该方法的预测稳定性较好,其训练、验证和测试集的均方根误差分别为 0.015 50、0.020 23 和 0.019 73。

关键词:城市固废焚烧;二噁英软测量;Bagging 半监督;伪标记样本;随机森林;深度森林回归
中图分类号:TH89 **文献标识码**:A **国家标准学科分类代码**:510.8060

Soft sensor of dioxin emission concentration based on Bagging semi-supervised deep forest regression

Xu Wen, Tang Jian, Xia Heng, Qiao Junfei
(*Faculty of Information Technology, Beijing University of Technology, Beijing 100124, China*)

Abstract: Dioxin (DXN), known as “the poison of the century”, is one of the by-products emitted by the municipal solid waste incineration (MSWI) process. Limited by the technical difficulty, time and economic cost of DXN emission concentration detection, it is difficult to obtain sufficient labeled samples for building a DXN emission soft sensor model. To effectively utilize a large number of unlabeled samples collected by the field control system and solve the problem of poor generalization performance of traditional shallow learning models, a soft-sensor method of DXN emission concentration based on Bagging semi-supervised deep forest regression (DFR) is proposed. First, multiple training subsets are obtained by resampling the original labeled dataset based on the Bagging mechanism, and multiple random forest (RF) models with diversities are formulated. Then, the RF model is iteratively updated, the nearest neighbor set is selected and the generalization performance strategies are evaluated, which are all used to obtain high-confidence pseudo-labeled samples. Finally, a DFR model is constructed based on the pseudo-labeled and original labeled sample sets. The effectiveness of the proposed method is evaluated with the actual DXN detection data of MSWI power plant in Beijing. It shows that the propose method has well prediction stability, and the root mean square errors are 0.015 50, 0.020 23 and 0.019 73 for training, validation and testing datasets respectively.

Keywords: municipal solid waste incineration; soft sensor of dioxin; Bagging semi-supervised; pseudo-labeled samples; random forest; deep forest regression

收稿日期:2022-03-07 Received Date: 2020-03-07
*基金项目:北京市自然科学基金(4212032)、国家自然科学基金(62073006)项目资助

0 引 言

城市固废焚烧(municipal solid waste incineration, MSWI)技术相对于填埋、生物处理等传统处理方法具有无害化、减量化、资源化等显著优势[1]。然而,该工艺过程排放的副产品之一是对生态环境和人类健康具有极大危害的持久性污染物二噁英(dioxin, DXN)[2]。实时检测其排放浓度对 MSWI 过程的运行优化和城市环境污染控制具有重要意义[3]。但由于 DXN 排放浓度检测的技术难度以及高经济与时间成本,用于构建其预测模型的样本数量稀少[4]。因此,面向 MSWI 过程的 DXN 排放浓度软测量建模存在有标记样本稀缺的问题。同时,现场控制系统可采集的大量无标记样本却未能被有效利用。针对这一问题,能够有效利用大量无标记样本构建软测量模型的半监督学习策略在类似行业获得成功应用[5-6]。

半监督学习的本质是通过获得无标记样本伪标签的方式扩展数量有限的原始标记样本,进而提高学习器的泛化性能,其关键是如何评估伪标记样本的置信度[7]。因此,当存在大量无标记样本时,需要对无标记样本进行筛选以确认其是否对所构建的软测量模型具有正向作用。Kang 等[8]建立了基于自训练策略的半监督支持向量回归(SS-SVR)模型,采用对无标记样本及其伪标签进行重采样的方式扩展训练数据集。史旭东等[9]建立了基于改进自训练算法的高斯过程回归软测量模型,先利用相似度估计无标记样本缺失的主导变量值后再筛选估计样本,进而将泛化能力强的伪标记样本选入原始标记样本集后建立软测量模型。但上述方法存在的弊端是:若选取了与有标记样本差异较大的无标记样本进行标记,会导致模型无法修正自训练过程的累计偏差。同时,上述方法在标记样本过程中采用传统单学习器构建软测量模型,其泛化性能有待提升。

研究表明,能够综合多个学习器的集成学习机制能够有效提高软测量模型的泛化性能,并进一步降低半监督学习的泛化误差[10]。Bagging 作为一种有效的并行集成机制已广泛应用于半监督领域[11-12]。随机森林(random forest, RF)是基于 Bagging 和随机子空间(random subspace method, RSM)的多决策树(decision tree, DT)集成学习算法[13],在面对小样本高维数据时具有良好的泛化性能,并且对数据中存在的噪声和异常值具有高包容性[14]。目前,RF 已广泛应用于医学图像分类与人脸识别、故障诊断、数据异常检测和关键参数预测等领域[15-19]。同时,许多学者提出了半监督 RF 方法,如 Leistner 等[20]将无标记样本的伪标签作为优化变量构建 RF 模型;Lu 等[21]基于信息熵协同训练半监督 RF 模型评估抑郁症状严重程度。但是,上述研究并未充分考虑伪标记初始模型的多样性,以及如何结合 Bagging 机制和 RF 获取更为有效的伪标记样本。

目前,深度学习(deep learning, DL)以其强大的特征表征能力和端到端的学习机制在多个领域获得成功应用[22]。基于神经网络的传统 DL 具有模型可解释性差、样本需求量大、超参数难调等缺点[23]。为有效结合深度学习和 RF 的优势,Kontschieder 等[24]提出了深度神经决策森林(deep neural decision forests, DNDF)算法。进一步,Zhou 等[25]提出了由多粒度扫描和级联森林组成的深度森林(deep forest, DF)算法用于图像分类问题,研究表明其在建模样本需求数量、超参数调节、泛化性能等方面均具有优势。针对难测参数的回归建模问题,汤等提出了深度森林回归(deep forest regression, DFR)及其改进算法[26-27],但仍存在未有效利用大量无标记数据等问题。

基于上述分析,本文提出了基于 Bagging 半监督 DFR 的软测量建模方法。首先,基于 Bagging 机制以重采样原始标记数据集的方式获得多个训练子集,并构建具有差异性的多个 RF 模型;接着,采用迭代机制更新 RF 模型、获取近邻集合、性能评估策略获得高置信度伪标记样本;最后,基于伪标记和原始标记样本集构建 DFR 预测模型。采用实际工业数据验证了所提方法的有效性。

1 面向 DXN 半监督建模的 MSWI 过程描述

MSWI 过程中的 DXN 排放问题在 1977 年首次引起了研究人员的注意[28]。典型的 MSWI 工艺流程,包含固废储运、固废焚烧、余热回收、烟气处理和烟气排放 5 个部分。MSWI 过程包括 DXN 的产生、吸收和排放 3 个阶段,在固废焚烧阶段,为保证有机物的有效分解,通常要求焚烧炉内的烟气温度达到 850℃并至少保持 2 s;在烟气处理阶段,石灰和活性炭被喷射进入反应器中以去除酸性气体、吸附 DXN 和一些重金属物,使得烟气 G1 中的 DXN 被分为两部分。一部分被吸附进入飞灰储仓,另一部分经袋式过滤器后保留在烟气 G2 中,通过引风机排入烟囱后作为烟气 G3 排入大气。因此,G3 烟气处的 DXN 排放浓度与固废焚烧、烟气处理和烟气排放阶段的众多过程变量有关。

以 $\boldsymbol{X}_{\mathrm{MSWI}}$ 表示 MSWI 过程变量,半监督学习就是利用向有标记样本中加入更多无标记样本 $\boldsymbol{X}_{\mathrm{unlabeled}} \in \boldsymbol{X}_{\mathrm{MSWI}}$ 的伪标签辅助建模,进而获得更精确的 DXN 排放浓度预测值。

2 建模策略与算法实现

2.1 建模策略

本文提出基于 Bagging 半监督 DFR 的建模策略,其包含初始 RF 模型构建模块、伪标记样本获取模块和 DFR 预测模块,其策略图如图 1 所示。

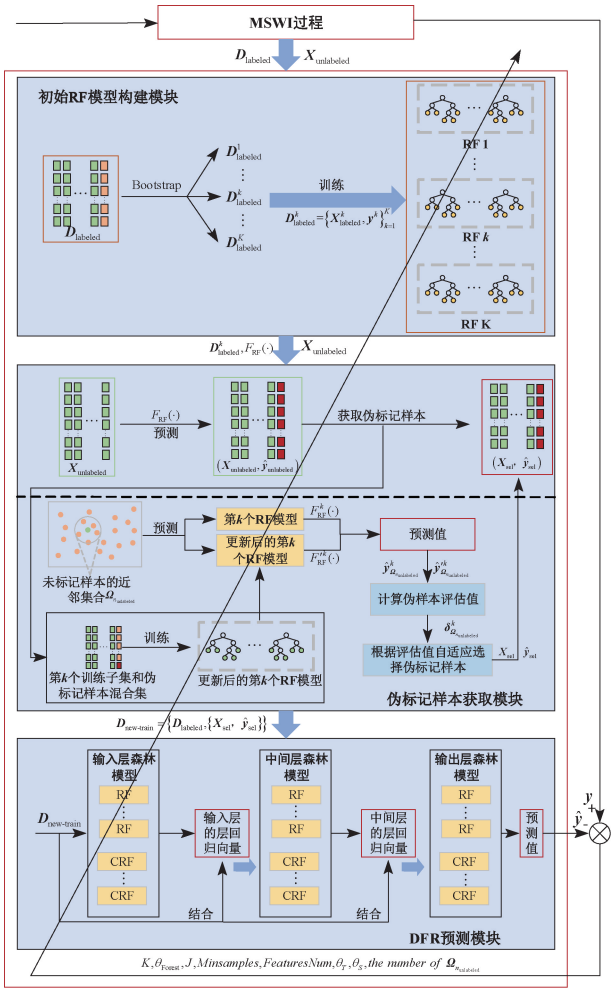


图 1　建模策略图

Fig. 1　Modeling strategy diagram

所提策略流程为:首先,基于原始标记样本集 $\boldsymbol{D}_{\text{labeled}}$ 随机采样得到的多个子集,并构建多个 RF 模型;接着,对无标记样本 $\boldsymbol{X}_{\text{unlabeled}}$ 进行伪标记,并选择具有高置信度的伪标记样本;最后,利用混合数据集 $\boldsymbol{D}_{\text{new-train}}$ 构建 DFR 预测模型。

2.2　算法实现

1)初始 RF 模型构建模块

原始标记样本集记为 $\boldsymbol{D}_{\text{labeled}}=\{\boldsymbol{x}_{n_{\text{labeled}}},\boldsymbol{y}_{n_{\text{labeled}}}\}_{n_{\text{labeled}}=1}^{N_{\text{labeled}}}$。

相应地,从输入特征的视角,第 n_{labeled} 个样本表示为下式:

$$\boldsymbol{x}_{n_{\text{labeled}}}=[x_{n_{\text{labeled}},1},x_{n_{\text{labeled}},2},\cdots,x_{n_{\text{labeled}},M}] \tag{1}$$

其中,N_{labeled} 和 M 分别为样本数和输入变量维数。

首先,通过有放回的抽样方法对 $\boldsymbol{D}_{\text{labeled}}$ 进行采样,进而得到 K 个训练子集。此处,将第 k 个训练子集记为 $\boldsymbol{D}_{\text{labeled}}^{k}$,进而全部训练子集可表示为:

$$\{\boldsymbol{D}_{\text{labeled}}^{k}\}_{k=1}^{K}=f_{\text{sam}}(\boldsymbol{D}_{\text{labeled}},N_{\text{labeled}},K)=\{\{\boldsymbol{x}_{n_{\text{labeled}}}^{k},y_{n_{\text{labeled}}}^{k}\}_{n_{\text{labeled}}=1}^{N_{\text{labeled}}}\}_{k=1}^{K}=\{\boldsymbol{X}_{\text{labeled}}^{k},\boldsymbol{y}_{\text{labeled}}^{k}\}_{k=1}^{K} \tag{2}$$

其中,$f_{\text{sam}}(\cdot)$ 表示样本抽样函数,$\boldsymbol{X}_{\text{labeled}}^{k}$ 和 $\boldsymbol{y}_{\text{labeled}}^{k}$ 表示第 k 个训练子集的输入和输出。

接着,基于训练子集构建 RF 模型。此处以 $\boldsymbol{D}_{\text{labeled}}^{k}$ 为例描述构建过程,采用结合 $f_{\text{sam}}(\cdot)$ 和 RSM 的方法对 $\boldsymbol{D}_{\text{labeled}}^{k}$ 进行共计 J 次的样本和特征随机采样,第 j 次产生子集 $\boldsymbol{D}_{\text{labeled}}^{k,j}$ 的过程如下式:

$$\boldsymbol{D}_{\text{labeled}}^{k,j}=f_{\text{RSM}}(f_{\text{sam}}(\boldsymbol{D}_{\text{labeled}}^{k},N_{\text{labeled}}),J,M_{\text{labeled}}^{k,j})=\{\boldsymbol{x}_{n_{\text{labeled}}}^{k,j},y_{n_{\text{labeled}}}^{k,j}\}_{n_{\text{labeled}}=1}^{N_{\text{labeled}}} \tag{3}$$

其中,$f_{\text{RSM}}(\cdot)$ 表示用于采样特征的 RSM 函数;$M_{\text{labeled}}^{k,j}$ 表示子训练集所选择的特征个数,通常存在 $M_{\text{labeled}}^{k,j}\ll M$。

在子集 $\boldsymbol{D}_{\text{labeled}}^{k,j}$ 所在的空间中将每个区域划分为两个子区域 R_1 和 R_2,并在每个子区域上构建 DT。遍历全部样本和特征,寻找最优变量编号和切分点取值 $(M_{\text{labeled,sel}}^{k,j},M_{\text{S}})$ 的过程为求解如下优化问题:

$$\begin{cases}(M_{\text{labeled,sel}}^{k,j},M_{\text{S}})=\min\Big[\sum^{R_1}((y_{n_{\text{labeled}}}^{k,j})_{R_1}-(\bar{y}_{\text{labeled}}^{k})_{R_1})^2+\sum^{R_2}((y_{n_{\text{labeled}}}^{k,j})_{R_2}-(\bar{y}_{\text{labeled}}^{k})_{R_2})^2\Big]\\ \text{s.t.}\begin{cases}R_1>\theta_{\text{Forest}}\\ R_2>\theta_{\text{Forest}}\end{cases}\end{cases} \tag{4}$$

其中,$(y_{n_{\text{labeled}}}^{k,j})_{R_1}$ 和 $(y_{n_{\text{labeled}}}^{k,j})_{R_2}$ 表示在 R_1 和 R_2 区域的某个测量值;$(\bar{y}_{\text{labeled}}^{k})_{R_1}$ 和 $(\bar{y}_{\text{labeled}}^{k})_{R_2}$ 表示 R_1 和 R_2 区域中全部测量值的平均值;θ_{Forest} 表示叶节点包含的样本数量阈值。

通过求解上式,优选得到的 $(M_{\text{labeled,sel}}^{k,j},M_{\text{S}})$ 用于划分 $\boldsymbol{D}_{\text{labeled}}^{k,j}$ 区域和确定相应的输出值,准则如下:

$$\begin{cases}R_1(M_{\text{labeled,sel}}^{k,j},M_{\text{S}})=\{\boldsymbol{x}_{n_{\text{labeled}}}^{k,j}\mid\boldsymbol{x}_{n_{\text{labeled}}}^{k,j}\leqslant M_{\text{S}}\}\\ R_2(M_{\text{labeled,sel}}^{k,j},M_{\text{S}})=\{\boldsymbol{x}_{n_{\text{labeled}}}^{k,j}\mid\boldsymbol{x}_{n_{\text{labeled}}}^{k,j}>M_{\text{S}}\}\end{cases} \tag{5}$$

进一步,对两个子区域重复上述步骤,直到叶节点中的样本数小于设定的阈值 θ_{Forest}。进而,将输入空间划分为 R_r 个区域,将获得的 DT 模型记为下式:

$$\Gamma^{j}(\cdot)=\sum_{r=1}^{R}c_{p,M_{\text{labeled}}^{k,j}}^{r}I(\boldsymbol{x}_{n_{\text{labeled}}}^{k,j}\in R_r) \tag{6}$$

$$c_{p,M_{\text{labeled}}^{k,j}}^{r}=\frac{1}{N_{R_r\boldsymbol{x}_{n_{\text{labeled}}}^{k,j}}}\sum_{\in R_r(M_{\text{labeled}}^{k,j},M_{\text{S}})}y_{\text{labeled},R_r}^{j,i} \tag{7}$$

其中,N_{R_r} 表示 R_r 区域内所包含样本个数;$y_{\text{labeled},R_r}^{j,i}$ 表示 R_r 区域内第 j 个子集的第 i 个真值;$I(\cdot)$ 为指示函数,即当 $\boldsymbol{x}_{n_{\text{labeled}}}^{k,j}\in R_r$ 存在时 $I(\cdot)=1$,否则 $I(\cdot)=0$。

在 $\boldsymbol{D}^k_{\text{labeled}}$ 上重复上述过程 J 次，得到的 RF 模型可表示如下：

$$F^k_{\text{RF}}(\cdot)=\frac{1}{J}\sum_{j=1}^{J}\Gamma^j(\cdot) \tag{8}$$

进一步，在 $\boldsymbol{D}_{\text{labeled}}$ 上重复 K 次，即得到 K 个初始 RF 模型，如下：

$$F_{\text{RF}}(\cdot)=\{F^k_{\text{RF}}(\cdot)\}^K_{k=1} \tag{9}$$

2）伪标记样本获取模块

（1）更新初始 RF 模型

无标记样本 $\boldsymbol{X}_{\text{unlabeled}}=[\boldsymbol{x}_1,\cdots,\boldsymbol{x}_{n_{\text{unlabeled}}},\cdots,\boldsymbol{x}_{N_{\text{unlabeled}}}]^{\text{T}}$ 中，第 $n_{\text{unlabeled}}$ 个样本 $\boldsymbol{x}_{n_{\text{unlabeled}}}$ 可表示为：

$$\boldsymbol{x}_{n_{\text{unlabeled}}}=[x_{n_{\text{unlabeled}},1},x_{n_{\text{unlabeled}},2},\cdots,x_{n_{\text{unlabeled}},M}] \tag{10}$$

以第 k 个初始 RF 模型 $F^k_{\text{RF}}(\cdot)$ 为例。首先，基于初始模型获得无标记样本 $\boldsymbol{x}_{n_{unlabeled}}$ 的伪标签，如下：

$$\hat{y}_{\boldsymbol{x}_{n_{\text{unlabeled}}}}=F^k_{\text{RF}}(\boldsymbol{x}_{n_{\text{unlabeled}}}) \tag{11}$$

接着，将由伪标签和 $\boldsymbol{D}^k_{\text{labeled}}$ 组合得到的新数据集 $\boldsymbol{D}'^k_{\text{labeled}}$ 表示为：

$$\boldsymbol{D}'^k_{\text{labeled}}=\{\boldsymbol{D}^k_{\text{labeled}},\{\boldsymbol{x}_{n_{\text{unlabeled}}},\hat{y}_{\boldsymbol{x}_{n_{\text{unlabeled}}}}\}\} \tag{12}$$

最后，基于 $\boldsymbol{D}'^k_{\text{labeled}}$ 构建更新的 RF 模型 $F'^k_{\text{RF}}(\cdot)$，构建过程如 2.2 节 1）中所示。

（2）获取无标记样本的近邻集合

在获得采用伪标记样本更新的 RF 模型 $F'^k_{\text{RF}}(\cdot)$ 后，需要判断该伪标记样本是否能够提升 $F'^k_{\text{RF}}(\cdot)$ 的性能。在本文中，通过计算有标记样本中无标记样本近邻集均方根误差（root mean square error，RMSE）的方式进行评估。

以 $\boldsymbol{x}_{n_{\text{unlabeled}}}$ 为例，其在标记子集 $\boldsymbol{D}^k_{\text{labeled}}$ 中的近邻集由距离向量确定，如下：

$$\boldsymbol{L}^k=[L^k_1,\cdots,L^k_{n_{\text{labeled}}},\cdots,L^k_{N_{\text{labeled}}}] \tag{13}$$

其中，$L^k_{n_{\text{labeled}}}$ 代表无标记样本 $\boldsymbol{x}_{n_{\text{unlabeled}}}$ 与第 n_{labeled} 个标记样本间的距离值，计算如下：

$$L^k_{n_{\text{labeled}}}=\|\boldsymbol{x}_{n_{\text{unlabeled}}}-\boldsymbol{x}^k_{n_{\text{labeled}}}\|_2=\sum_{m=1}^{M}\sqrt{(x_{n_{\text{unlabeled}},m}-x^k_{n_{\text{labeled}},m})^2} \tag{14}$$

在获得 $\boldsymbol{L}^k$ 后，依据数值大小进行升序排列，排序后的向量记为：

$$\boldsymbol{L}^k_{\text{Sort}}=f_{\text{Sort}}(\boldsymbol{L}^k)=f_{\text{Sort}}(L^k_1,L^k_2,\cdots,L^k_{N_{\text{labeled}}})=[L^k_{\text{Sort},1},\boldsymbol{L}^k_{\text{Sort},2},\cdots,L^k_{\text{Sort},N_{\text{labeled}}}]^{\text{T}} \tag{15}$$

其中，$f_{\text{Sort}}(\cdot)$ 为排序函数。

此外，设定近邻集的数量阈值为 N_{near}，进而得到 $\boldsymbol{L}^k_{\text{Sel}}$，记为：

$$\boldsymbol{L}^k_{\text{Sel}}=[L^k_{\text{Sort},1},L^k_{\text{Sort},2},\cdots,L^k_{\text{Sort},N_{\text{near}}}]^{\text{T}} \tag{16}$$

进一步，得到 N_{near} 个标记样本组成的近邻集合，如下：

$$\boldsymbol{\Omega}_{n_{\text{unlabeled}}}=\{\boldsymbol{x}^k_{\text{Sel},n_{\text{near}}}\}^{N_{\text{near}}}_{n_{\text{near}}=1} \tag{17}$$

（3）评估伪标记样本

首先，基于 $F^k_{\text{RF}}(\cdot)$ 和 $F'^k_{\text{RF}}(\cdot)$ 计算得到近邻集合 $\boldsymbol{\Omega}_{n_{\text{unlabeled}}}$ 的预测值 $\hat{\boldsymbol{y}}^k_{\boldsymbol{\Omega}_{n_{\text{unlabeled}}}}=F^k_{\text{RF}}(\boldsymbol{\Omega}_{n_{\text{unlabeled}}})$ 和 $\hat{\boldsymbol{y}}'^k_{\boldsymbol{\Omega}_{n_{\text{unlabeled}}}}=F'^k_{\text{RF}}(\boldsymbol{\Omega}_{n_{\text{unlabeled}}})$。

采用下式评估加入伪标记样本后对近邻集合的预测性能：

$$\delta^k_{\boldsymbol{\Omega}_{n_{\text{unlabeled}}}}=\sum_{\boldsymbol{x}_{\boldsymbol{\Omega}}\in\boldsymbol{\Omega}_{n_{\text{unlabeled}}}}\left(\begin{array}{l}(\boldsymbol{y}^k_{\boldsymbol{\Omega}_{n_{\text{unlabeled}}}}-\hat{\boldsymbol{y}}^k_{\boldsymbol{\Omega}_{n_{\text{unlabeled}}}})^2-\\(\boldsymbol{y}^k_{\boldsymbol{\Omega}_{n_{\text{unlabeled}}}}-\hat{\boldsymbol{y}}'^k_{\boldsymbol{\Omega}_{n_{\text{unlabeled}}}})^2\end{array}\right) \tag{18}$$

其中，$\delta^k_{\boldsymbol{\Omega}_{n_{\text{unlabeled}}}}$ 称为伪样本评估值；$\boldsymbol{y}^k_{\boldsymbol{\Omega}_{n_{\text{unlabeled}}}}$ 为近邻样本 $\boldsymbol{x}_{\boldsymbol{\Omega}_{n_{\text{unlabeled}}}}$ 的真值。

显然，$\delta^k_{\boldsymbol{\Omega}_{n_{\text{unlabeled}}}}$ 的值越大，表示伪标记样本（$\boldsymbol{x}_{n_{\text{unlabeled}}}$，$\hat{y}_{\boldsymbol{x}_{n_{\text{unlabeled}}}}$）对提高模型性能的积极影响越大。

上述过程描述仅针对于单个无标记样本进行评估。为优选无标记样本，此处将选择无标记样本的数量记为 θ_s。重复上述针对单个无标记样本的评估过程，在无标记样本数量达到 θ_s 个后，选择最大的伪样本评估值 $\delta^{k,\max}_{\boldsymbol{\Omega}_{n_{\text{unlabeled}}}}$，并判断其是否大于 0；若大于 0 则选择 $\delta^{k,\max}_{\boldsymbol{\Omega}_{n_{\text{unlabeled}}}}$ 对应的伪标记样本，若小于 0 则进入下次迭代。

为保证 $\boldsymbol{D}^k_{\text{labeled}}$ 能够得到一定数量的伪标记样本，将上述迭代过程的阈值记为 θ_T，并针对 $\boldsymbol{D}^k_{\text{labeled}}$ 在迭代过程获得的伪标记样本进行存储。将通过上述过程获得 $\boldsymbol{D}^k_{\text{labeled}}$ 标记的伪标记样本记为 $\{\boldsymbol{X}^k_{\text{sel}},\hat{\boldsymbol{y}}^k_{\text{sel}}\}$。

重复上述过程 K 次，获得的伪标记样本集记为：

$$\{\boldsymbol{X}_{\text{sel}},\hat{\boldsymbol{y}}_{\text{sel}}\}=\left\{\begin{array}{c}\{\boldsymbol{X}^1_{\text{sel}},\hat{\boldsymbol{y}}^1_{\text{sel}}\}\\\vdots\\\{\boldsymbol{X}^K_{\text{sel}},\hat{\boldsymbol{y}}^K_{\text{sel}}\}\end{array}\right\} \tag{19}$$

3）DFR 预测模块

基于混合样本集 $\boldsymbol{D}_{\text{new-train}}=\{\boldsymbol{D}_{\text{labeled}},\{\boldsymbol{X}_{\text{sel}},\hat{\boldsymbol{y}}_{\text{sel}}\}\}=\{\boldsymbol{X}_{\text{new}},\boldsymbol{y}_{\text{new}}\}$ 构建 DFR 模型，后者以 Stacking 方式组合多个不同类别的森林模型，包含输入层、中间层和输出层森林模型。

针对输入层森林模型，随机采样 $\boldsymbol{D}_{\text{new-train}}$，构建基于 RF 和完全 RF（CRF）的子森林模型，其第 i 个子森林模型的预测均值通过下式计算得到：

$$\hat{y}_{1,i}=\frac{1}{J}\sum_{j=1}^{J}\hat{\boldsymbol{y}}^j_{1,i},i=1,2,\cdots,I \tag{20}$$

其中，I 为子森林数量。接着，通过选择 k_{kNN} 个邻近 $\hat{\bar{y}}_{1,i}$ 的预测值形成回归向量 $\hat{\boldsymbol{y}}^{k_{\text{kNN}}}_{1,i}$。重复 I 次后得到层回归向量 $\hat{\boldsymbol{y}}^{\text{regvec}}_1=\{\hat{\boldsymbol{y}}^{\text{kNN}}_{1,i}\}^I_{i=1}$，将其与原始输入特征向量 $\boldsymbol{X}_{\text{new}}$ 组合得到中间层森林模型（第 2 层）的输入，即增强回归向

量$\begin{bmatrix}\hat{\boldsymbol{y}}_1^{\text{regvec}}\\ \boldsymbol{X}_{\text{new}}\end{bmatrix}^{\text{T}}$。

针对中间层森林模型,第 λ 层的输入可表示为:

$$\boldsymbol{D}_\lambda=\left\{\begin{bmatrix}\hat{\boldsymbol{y}}_{\lambda-1}^{\text{regvec}}\\ \boldsymbol{X}_{\text{new}}\end{bmatrix}^{\text{T}},\boldsymbol{y}_{\text{new}}\right\},\lambda=2,3,\cdots,L-1 \tag{21}$$

相应地,其子森林模型的预测均值输出为:

$$\hat{\bar{y}}_{\lambda,i}=\frac{1}{J}\sum_{j=1}^{J}\hat{y}_{\lambda,i}^{j} \tag{22}$$

其中,$\begin{bmatrix}\hat{\boldsymbol{y}}_{\lambda-1}^{\text{regvec}}\\ \boldsymbol{X}_{\text{new}}\end{bmatrix}^{\text{T}}$ 为第 $\lambda-1$ 层森林模型的层回归向量 $\hat{\boldsymbol{y}}_{\lambda-1}^{\text{regvec}}$ 和原始特征向量 $\boldsymbol{X}_{\text{new}}$ 组成的增强回归向量;$\hat{y}_{\lambda,i}^{j}$ 由第 λ 个森林模型中的第 i 层子森林中的 J 个 DT 模型生成。类似输入层,选择 k_{kNN} 个邻近 $\hat{\bar{y}}_{\lambda,i}$ 的预测值以获得第 i 个子森林的回归向量 $\hat{\boldsymbol{y}}_{\lambda,i}^{k_{\text{kNN}}}$,进而得到第 λ 个森林模型的层回归向量 $\hat{\boldsymbol{y}}_{\lambda}^{\text{regvec}}=\{\hat{\boldsymbol{y}}_{\lambda,i}^{k_{\text{kNN}}}\}_{i=1}^{I}$。

针对输出层森林模型,即第 L 层森林模型,其训练数据集为 $\boldsymbol{D}_L=\left\{\begin{bmatrix}\hat{\boldsymbol{y}}_{L-1}^{\text{regvec}}\\ \boldsymbol{X}_{\text{new}}\end{bmatrix}^{\text{T}},\boldsymbol{y}_{\text{new}}\right\}$。相应地,第 i 个子森林模型的预测均值由下式计算:

$$\hat{\bar{y}}_{L,i}=\frac{1}{J}\sum_{j=1}^{J}\hat{y}_{L,i}^{j} \tag{23}$$

其中,$\hat{y}_{L,i}^{j}$ 为第 L 层中第 i 个子森林模型的每个 DT 模型生成的预测值。重复上述步骤 I 次,即可得到 I 个子森林模型的预测输出集 $\{\hat{\boldsymbol{y}}_{L,i}^{k_{\text{kNN}}}\}_{i=1}^{I}$。进一步,计算 I 个子森林模型预测值的平均值获得 DFR 模型的最终预测值为:

$$\hat{y}=\frac{1}{I}\sum_{i=1}^{I}\hat{y}_{L,i} \tag{24}$$

3　实验验证

本文采用式(18)评价伪标记样本的置信度高低;采用 RMSE 和平均绝对误差(mean absolute error, MAE)评价拟合性能,公式如下:

$$RMSE=\sqrt{\frac{1}{N}\sum_{n=1}^{N}(\hat{y}^n-y^n)^2} \tag{25}$$

$$MAE=\frac{1}{N}\sum_{n=1}^{N}|\hat{y}^n-y^n| \tag{26}$$

3.1　数据描述

本节采用的建模数据源自北京某 MSWI 电厂 2#炉 2012~2018 年间 DXN 排放浓度检测当天及前后 3 天的过程数据,包括:74 个有标记样本,前后 3 天的 436 个无标记样本。其中,有标记样本作为训练集、验证集和测试集的来源,划分比例为 2:1:1。

3.2　实验结果

1)初始 RF 模型构建结果

基于 Bagging 机制获得多个训练子集以构建具有差异性的 RF 模型,表 1 描述了 DXN 数据集初始 RF 模型构建过程的参数设置情况。

表 1　初始 RF 模型构建参数

Table 1　Initial RF model construction parameters

数据集	训练子集个数(K)	训练集样本个数	验证集样本个数	测试集样本个数	决策树个数(J)	最小样本个数(*Minsample*)	选取特征数量(*Features Num*)
DXN	30	38	18	18	50	8	13

2)伪标记样本获取结果

本文设定子训练集的个数 $K=30$,每个训练集在迭代次数 $\theta_T=10$ 的过程中选择不同的伪标记样本。图 2 为 DXN 数据集在一次迭代过程中的 2 个训练子集选取伪标记样本的过程,其中:横坐标代表不同的无标记样本,纵坐标为伪标记样本评估值,设置阈值为 0。

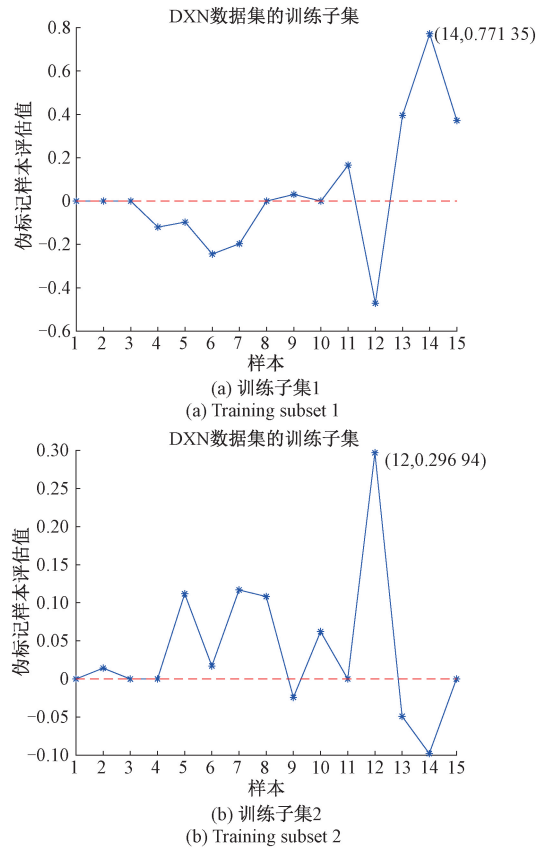


图 2　选取伪标记样本过程

Fig. 2　Process of selecting the pseudo-labeled samples

在每次迭代过程中,各训练子集对选取出的无标记样本进行伪标记,按照式(18)计算伪样本评估值,选择最高置信度且大于阈值的伪标记样本并添加至最终的训练集中,最终伪标记样本为 298 个。

3) DFR 预测及比较结果

基于混合样本构建 DFR 模型预测 DXN 排放浓度,参数设置如下: $Minsamples = 8$, $FeaturesNum = 13$, $J = 50$, $\theta_T = 10$,无标记样本数 $\theta_S = 15$,近邻域集的样本个数为 30。同时,本文方法与 DT、半监督协同训练决策树(Co-DT)和 RF 进行对比,该 4 种方法分别运行 20 次,训练、验证和测试集 RMSE 和 MAE 结果的均值和方差如表 2 所示。

表 2 DXN 数据集的实验结果

Table 2 Experimental results on the DXN dataset

评价指标	方法	训练集		验证集		测试集	
		平均值	方差	平均值	方差	平均值	方差
RMSE	DT	$2.289\,4\times10^{-2}$	$2.523\,2\times10^{-5}$	$2.696\,6\times10^{-2}$	$4.796\,1\times10^{-5}$	$2.612\,4\times10^{-2}$	$4.384\,4\times10^{-5}$
	Co-DT	$2.186\,6\times10^{-2}$	$2.254\,9\times10^{-5}$	$2.552\,2\times10^{-2}$	$1.940\,8\times10^{-5}$	$2.365\,7\times10^{-2}$	$1.669\,2\times10^{-5}$
	RF	$1.704\,4\times10^{-2}$	$2.940\,3\times10^{-7}$	$2.049\,1\times10^{-2}$	$8.893\,1\times10^{-7}$	$1.976\,9\times10^{-2}$	$1.643\,3\times10^{-6}$
	本文	**$1.549\,6\times10^{-2}$**	**$1.902\,3\times10^{-7}$**	**$2.023\,0\times10^{-2}$**	**$4.817\,9\times10^{-7}$**	**$1.972\,9\times10^{-2}$**	**$5.709\,0\times10^{-7}$**
MAE	DT	$1.608\,3\times10^{-2}$	$1.576\,8\times10^{-5}$	$2.282\,0\times10^{-2}$	$4.572\,0\times10^{-5}$	$2.039\,1\times10^{-2}$	$2.644\,4\times10^{-5}$
	Co-DT	$1.742\,1\times10^{-2}$	$2.384\,6\times10^{-5}$	$2.168\,2\times10^{-2}$	$1.397\,8\times10^{-5}$	$2.011\,1\times10^{-2}$	$1.391\,0\times10^{-5}$
	RF	$1.389\,4\times10^{-2}$	**$2.399\,9\times10^{-7}$**	$1.946\,9\times10^{-2}$	$7.565\,0\times10^{-7}$	**$1.807\,5\times10^{-2}$**	$1.879\,8\times10^{-6}$
	本文	**$1.253\,5\times10^{-2}$**	$2.611\,1\times10^{-7}$	**$1.905\,0\times10^{-2}$**	**$5.315\,4\times10^{-7}$**	$1.832\,1\times10^{-2}$	**$7.015\,1\times10^{-7}$**

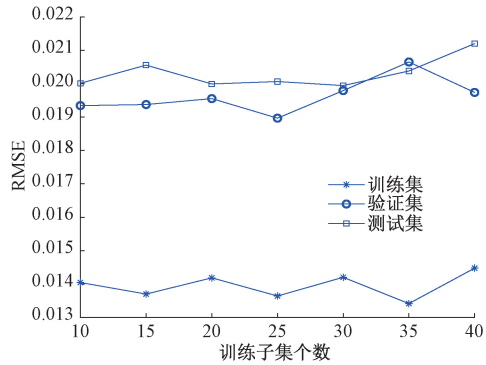
由表 2 可知,半监督 Co-DT 方法相较于 DT 方法得到的 RMSE 和 MAE 的平均值较小,这说明加入伪标记样本使得其预测精度优于仅使用有标记样本的 DT 算法,表明无标记样本可为建模过程提供更全面的数据信息;基于集成思想的 RF 算法的建模型精度优于使用单一学习器的 Co-DT 方法;本文方法结合了半监督学习和 Bagging 思想训练深度模型,在训练、验证和测试集上的 RMSE 平均值和方差均为最优,表明该模型性能优越,比其他方法稳定、泛化性能好;在 MAE 指标上的结果表明,验证集表现最佳,训练集和测试集的平均值虽最小但在方差上却略大于 RF 模型,说明该方法仍具有进一步的优化空间。

综上实验结果可知,加入一定数量的伪标记样本可提高软测量模型的泛化性能,验证了本文所提半监督策略的有效性。

3.3 参数分析

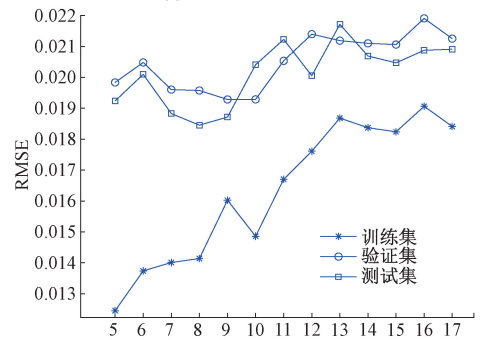
针对所提方法中的 3 个超参数,即 K、$Minsamples$ 和 $FeaturesNum$ 进行分析,这些参数与 RMSE 间的关系如图 3 所示。

由图 3 可知:1)子训练集的个数 K 对训练、验证和测试数据集的影响都较为平缓,综合考虑模型的训练时间,可以选择较为适中的值,本文中 $K = 25$ 较为合适;2)最小样本数量 $Minsamples$,对训练数据集的影响较大,呈现出其 RMSE 随着 $Minsamples$ 的增加而变大的趋势;但针对验证和测试样本而言,其 RMSE 呈现出先减小再增加最后趋于平缓的趋势,表明选择适合的 $Minsamples$ 是非常必要的;本文中 $Minsamples = 9$ 较为合适;3)特征数量 $FeaturesNum$,对训练集的影响并不如 $Minsamples$ 显著,但其 RMSE 也呈现出随 $FeaturesNum$ 的增加而降低的



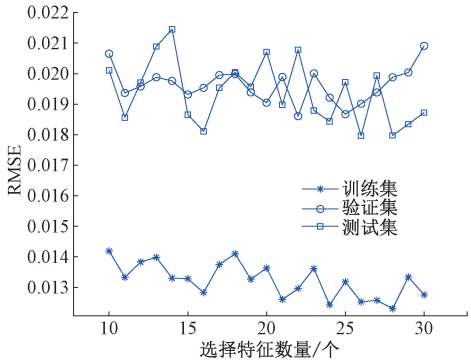
(a) 训练子集个数与RMSE的关系图

(a) Relation of K versus RMSE



(b) 最小样本个数与RMSE的关系图

(b) Relation of $Minsamples$ versus RMSE



(c) 选择特征数量与RMSE的关系图
(c) Relation of *FeaturesNum* versus RMSE

图 3　超参数与 RMSE 的关系图

Fig. 3　Relation of hyperparameters versus RMSE

趋势;针对验证和测试样本而言,RMSE 虽然呈现较大的波动,但总体趋势也是先降低再增加,并且测试数据呈现较大的波动性,表明建模数据质量还有待于进一步提高;本文中 *FeaturesNum* = 22 比较合适。

需要提出的是,上述针对超参数的分析所采用的是单因素方法。事实上,众多超参数之间还存在着相互影响的耦合作用,因此有必要对这些超参数进行同时的优化选择。

4　结　　论

为提高 DXN 软测量模型性能,充分利用 MSWI 过程的大量无标记过程数据,本文提出了一种面向 DXN 的半监督建模方法。主要贡献包括:1)提出基于 Bagging 半监督 DFR 的软测量模型框架,充分利用了半监督 RF 有效扩充标记样本和 DFR 能够提取深层特征的优势;2)提出基于 Bagging 随机抽样训练多个具有差异性 RF 模型,采用基于迭代机制的 RF 模型更新以及近邻集合选择和标记样本性能评估策略,自适应获得高置信度伪标记样本;3)基于工业数据得到的混合标记样本建立 DFR 软测量模型,能够有效提取混合样本的深度特征,验证了所提方法的有效性。进一步的研究工作包括:如何自适应设置阈值选择伪标记样本和如何对超参数进行全局的同时优化。

参考文献

[1] LI M F, ZHOU Y X, WANG G S, et al. Evaluation of atmospheric sources of PCDD/Fs, PCBs and PBDEs around an MSWI plant using active and passive air samplers[J]. Chemosphere, 2021:274.

[2] 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[J]. 自动化学报, 2020, 46(6): 1063-1089.
QIAO J F, GUO Z H, TANG J. Dioxin emission concentration measurement approaches for municipal solid wastes incineration process: A survey [J]. Acta Automatica Sinica, 2020,46(6): 1063-1089.

[3] 汤健, 王丹丹, 郭子豪, 等. 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[J]. 北京工业大学学报, 2021, 47(5): 431-443.
TANG J, WANG D D, GUO Z H, et al. Prediction of dioxin emission concentration in the municipal solid waste incineration process based on optimal selection of virtual samples[J]. Journal of Beijing University of Technology, 2021, 47(5): 431-443.

[4] XIA H, TANG J, ALJERF L. Dioxin emission prediction based on improved deep forest regression for municipal solid waste incineration process[J]. Chemosphere, 2022 (294): 133716.

[5] JIAN C X, YANG K J, AO Y H. Industrial fault diagnosis based on active learning and semi-supervised learning using small training set [J]. Engineering Applications of Artificial Intelligence, 2021, 104: 104365.

[6] 吕枫, 王义, 阮胡林, 等. 深度嵌入关系空间下齿轮箱标记样本扩充及其半监督故障诊断方法[J]. 仪器仪表学报, 2021, 42(2): 55-65.
LYU F, WANG Y, RUAN H L, et al. Gearbox labeled sample augmentation and its semi-supervised fault diagnosis method in deeply embedded relational space[J]. Chinese Journal of Scientific Instrument, 2021, 42(2): 55-65.

[7] HU Z, YANG Z, HU X, et al. SimPLE: Similar pseudo label exploitation for semi-supervised classification[C]. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021: 15094-15103.

[8] KANG P, KIM D, CHO S. Semi-supervised support vector regression based on self-training with label uncertainty: An application to virtual metrology in semiconductor manufacturing[J]. Expert Systems with Applications, 2016, 51: 85-106.

[9] 史旭东, 熊伟丽. 基于改进自训练算法的半监督 GPR 软测量建模[J]. 控制工程, 2020, 27(3): 451-455.

SHI X D, XIONG W L. Semi-supervised gaussian process regression modeling based on improved self-training algorithm[J]. Control Engineering of China, 2020, 27(3): 451-455.

[10] ZHOU Z H. When semi-supervised learning meets ensemble learning[J]. Frontiers of Electrical and Electronic Engineering in China, 2011, 6(1): 6-16.

[11] LI Y, SU L, CHEN J, et al. Semi-supervised question classification based on ensemble learning[C]. Advances in Swarm and Computational Intelligence, 2015: 341-348.

[12] HE X, JI J, LIU K X, et al. Soft sensing of silicon content via Bagging local semi-supervised models[J]. Sensors, 2019, 19(17): 3814-3814.

[13] BREIMAN L. Random forests[J]. Machine Learning, 2001, 45(1): 5-32.

[14] 方匡南, 吴见彬, 朱建平, 等. 随机森林方法研究综述[J]. 统计与信息论坛, 2011, 26(3): 32-38.
FANG K N, WU J B, ZHU J P, et al. A review of random forest method research[J]. Statistics & Information Forum, 2011, 26(3): 32-38.

[15] ARIF M. Classification of cardiotocograms using random forest classifier and selection of important features from cardiotocogram signal[J]. Biomaterials and Biomedical Engineering, 2015, 2(3): 173-183.

[16] 高学金, 马东阳, 韩华云, 等. 基于 DAE 和 TCN 的复杂工业过程故障预测[J]. 仪器仪表学报, 2021, 42(6): 140-151.
GAO X J, MA D Y, HAN H Y, et al. Fault prediction of complex industrial process based on DAE and TCN[J]. Chinese Journal of Scientific Instrument, 2021, 42(6): 140-151.

[17] DENG W, GUO Y, LIU J, et al. A missing power data filling method based on improved random forest algorithm[J]. Chinese Journal of Electrical Engineering, 2019, 5(4): 33-39.

[18] 王毅, 陈进, 李松浓, 等. 基于时频域分析和随机森林的故障电弧检测[J]. 电子测量与仪器学报, 2021, 35(5): 62-68.
WANG Y, CHEN J, LI S N, et al. Arc fault detection based on time and frequency analysis and random forest[J]. Journal of Electronic Measurement and Instrumentation, 2021, 35(5): 62-68.

[19] LI L, LIANG T C, AI S, et al. An improved random forest algorithm and its application to wind pressure prediction[J]. International Journal of Intelligent Systems, 2021, 36(8): 4016-4032.

[20] LEISTNER C, SAFFARI A, SANTNER J, et al. Semi-supervised random forests[C]. In IEEE Conference on Computer Vision, 2009: 506-513.

[21] LU S F, SHI X, LI M, et al. Semi-supervised random forest regression model based on co-training and grouping with information entropy for evaluation of depression symptoms severity[J]. Mathematical Biosciences and Engineering, 2021, 18(4): 4586-4602.

[22] TANG C W, YU C Y, GAO Y, et al. Deep learning in nuclear industry: A survey[J]. Big Data Mining and Analytics, 2022, 5(2): 140-160.

[23] 汤健, 夏恒, 乔俊飞, 等. 深度集成森林回归建模方法及应用[J]. 北京工业大学学报, 2021, 47(11): 1219-1229.
TANG J, XIA H, QIAO J F, et al. Deep ensemble forest regression modeling method with its application research[J]. Journal of Beijing University of Technology, 2021, 47(11): 1219-1229.

[24] KONTSCHIEDER P, FITERAU M, CRIMINISI A. Deep neural decision forests[C]. In IEEE International Conference on Computer Vision, 2015: 1467-1475.

[25] ZHOU Z H, FENG J. Deep forest[J]. National Science Review, 2019(6): 74-86.

[26] TANG J, XIA H, ZHANG J, et al. Deep forest regression based on cross-layer full connection[J]. Neural Computing and Applications, 2021, 33: 9307-9328.

[27] XIA H, TANG J, QIAO J, et al. DF classification algorithm for constructing a small sample size of data-oriented DF regression model[J]. Neural Computing and Applications, 2022, 34: 2785-2810.

[28] OLIE K, VERMEULEN P L, HUTZINGER O. Chlorodibenzo-p-dioxins and chlorodibenzofurans are trace components of fly ash and flue gas of some municipal incinerators in the netherlands[J]. Chemosphere, 1977, 6(8): 455-459.

作者简介

徐雯,2019 年于河南理工大学获得学士学位,现为北京工业大学硕士研究生,主要研究方向为基于半监督集成森林的城市固废过程二噁英智能预测。

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

Xu Wen received her B. Sc. degree from Henan Polytechnic University in 2019. She is currently a master student at Beijing University of Technology. Her main research interests include intelligent prediction of dioxins in municipal solid waste incineration process based on semi-supervised ensemble forest.

汤健(通信作者),2012 年于东北大学获得博士学位,现为北京工业大学教授,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。

E-mail:freeflytang@ bjut. edu. cn

Tang Jian (Corresponding author) received his Ph. D. degree from Northeastern University in 2012. He is currently a professor at Beijing University of Technology. His main research interests include small sample data modeling and intelligent control of municipal solid waste incineration process.

夏恒,2020 年于北京工业大学获得硕士学位,现为北京工业大学博士研究生,主要研究方向为小样本数据建模、城市固废处理过程的二噁英排放预测。

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

Xia Heng received his M. Sc. degree from Beijing University of Technology in 2020. He is currently a Ph. D. candidate at Beijing University of Technology. His main research interests include small sample data modeling and dioxin emission prediction in municipal solid waste incineration process.

乔俊飞,1998 年于东北大学获得博士学位,现为北京工业大学教授,主要研究方向为环保过程智能控制,神经网络结构设计与优化。

E-mail:junfeiq@ bjut. edu. cn

Qiao Junfei received his Ph. D. degree from Northeastern University in 1998. He is currently a professor at Beijing University of Technology. His main research interests include intelligent control of environmental protection processes, and neural network structure design and optimization.