(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 114398836 A
(43)申请公布日 2022.04.26

(21)申请号 202210059984.5

(22)申请日 2022.01.19

(71)申请人 北京工业大学
地址 100020 北京市朝阳区平乐园100号

(72)发明人 汤健 夏恒 璀璨麟 乔俊飞

(74)专利代理机构 北京鑫瑞森知识产权代理有限公司 11961
代理人 代芳

(51)Int.Cl.
*G06F 30/27*(2020.01)
*G06K 9/62*(2022.01)
*G06N 3/02*(2006.01)
*G06N 5/00*(2006.01)

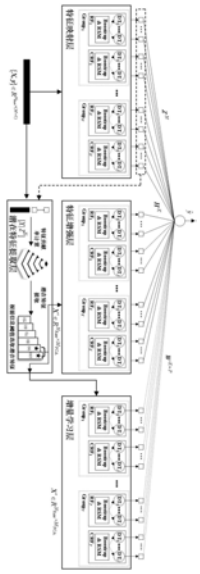
权利要求书6页 说明书17页 附图4页

(54)发明名称

基于宽度混合森林回归的MSWI过程二噁英排放软测量方法

(57)摘要

本发明提供了一种基于宽度混合森林回归的MSWI过程二噁英排放软测量方法,基于BLS框架,以非微分基学习器替换神经元构建面向小样本高维数据的BHFR软测量模型,BHFR软测量模型包括特征映射层、潜在特征提取层、特征增强层和增量学习层的构建:首先,构建由随机森林和完全随机森林组成的混合森林组进行高维特征映射;其次,依据贡献率对全联接混合矩阵的特征空间进行潜在特征提取,采用信息度量准则降低模型复杂度和计算消耗;然后,基于所提取潜在信息训练特征增强层以增强特征表征能力;最后,通过增量式学习策略构建增量学习层,采用Moore-Penrose伪逆获得权重矩阵,进而实现高精度建模。在高维基准数据集和工业过程DXN数据集上验证了所提方法的有效性和合理性。



CN 114398836 A

1.一种基于宽度混合森林回归的MSWI过程二噁英排放软测量方法，基于BLS框架，以非微分基学习器替换神经元构建面向小样本高维数据的BHFR软测量模型，其特征在于，所述BHFR软测量模型包括特征映射层、潜在特征提取层、特征增强层和增量学习层的构建，具体包括以下步骤：

S1，构建特征映射层，构建由随机森林RF和完全随机森林CRF组成的混合森林组对高维特征进行映射；

S2，构建潜在特征提取层，依据贡献率对全联接混合矩阵的特征空间进行潜在特征提取，基于信息度量准则保证潜在有价值信息的最大化传递和最小化冗余，降低模型复杂度和计算消耗；

S3，构建特征增强层，基于所提取的潜在特征训练特征增强层以进一步增强特征表征能力；

S4，构建增量学习层，通过增量式学习策略构建增量学习层，采用Moore-Penrose伪逆获得权重矩阵，进而实现BHFR软测量模型的高精度建模；

S5，采用高维基准数据集和工业过程DXN数据集验证所述软测量模型；

S6，采用步骤S1-S5建立的软测量模型，对MSWI过程二噁英排放进行软测量。

2.根据权利要求1所述的基于宽度混合森林回归的MSWI过程二噁英排放软测量方法，其特征在于，步骤S1，构建特征映射层，构建由随机森林RF和完全随机森林CRF组成的混合森林组对高维特征进行映射，具体包括：

设原始数据为{X,y}，其中 $\boldsymbol{X} \in R^{N_{\mathrm{Raw}} \times M}$ 是原始输入数据，$N_{Raw}$是原始数据的数量，M是原始输入数据的维数，其来源于MSWI过程的六个不同阶段，以秒为单位在DCS系统采集与存储，$\boldsymbol{y} \in R^{N_{\mathrm{Raw}} \times 1}$是DXN排放浓度的输出真值，其来源于采用离线检测法得到排放物DXN检测样本；以特征映射层的第nth个混合森林组为例描述特征映射层的建模过程：

对{X,y}进行Bootstrap和随机子空间RSM采样，获得混合森林组模型的J个训练子集，如下：

$$\left\{\boldsymbol{X}_{\mathrm{Bootstrap}}^{n,j}, \boldsymbol{y}_{\mathrm{Bootstrap}}^{n,j}\right\}_{j=1}^{J} = \varphi_n^{\mathrm{FML}}\left(\phi_n^{\mathrm{FML}}((\boldsymbol{X},\boldsymbol{y}), P_{\mathrm{Bootstrap}})\right) \tag{1}$$

其中，$\boldsymbol{X}_{\mathrm{Bootstrap}}^{n,j}$和$\boldsymbol{y}_{\mathrm{Bootstrap}}^{n,j}$为第J个训练子集的输入和输出，$\phi_n^{\mathrm{FML}}(\cdot)$和$\varphi_n^{\mathrm{FML}}(\cdot)$表示特征映射层中对第nth个混合森林组的Bootstrap和RSM采样，$P_{\mathrm{Bootstrap}}$表示Bootstrap采样概率；

基于$\left\{\boldsymbol{X}_{\mathrm{Bootstrap}}^{n,j}, \boldsymbol{y}_{\mathrm{Bootstrap}}^{n,j}\right\}_{j=1}^{J}$训练包含J个决策树的混合森林算法，其中特征映射层中的第nth个混合森林组的第jth个决策树表示如下：

$$f_{n,j}^{\mathrm{DT}}(\cdot) = \sum_{l=1}^{L} c_l I(\boldsymbol{x}_{\mathrm{Bootstrap}}^{n,j} \in R_l), \quad l = 1,2,...,L \tag{2}$$

其中，L表示决策树叶节点数量，I(•)表示指示函数，$c_1$采用递归分裂方式计算；

RF中决策树的分裂损失函数$\Omega_i(\cdot)$表示为：

$$\begin{aligned}\Omega_i(s,v) &= \min\left([\boldsymbol{y}_L - E[\boldsymbol{y}_L]] + [\boldsymbol{y}_R - E[\boldsymbol{y}_R]]\right) \\ &= \min\left(\sum_{\boldsymbol{x}_{\mathrm{Bootstrap}}^{n,j} \in R_{\mathrm{L}}} (y_{\mathrm{L}}^{\mathrm{i}} - c_{\mathrm{L}})^2 + \sum_{\boldsymbol{x}_{\mathrm{Bootstrap}}^{n,j} \in R_{\mathrm{R}}} (y_{\mathrm{R}}^{\mathrm{i}} - c_{\mathrm{R}})^2\right)\end{aligned} \tag{3}$$

其中，$\Omega_i(s,v)$表示第sth个特征的值v作为切分准则的损失函数值，$y_L$表示左叶节点的DXN排放浓度真值向量，$E[y_L]$表示$y_L$的数学期望，$y_R$表示右叶节点的DXN排放浓度真值向量，$E[y_R]$表示$y_R$的数学期望，$y_L^i$表示左叶节点第i个DXN排放浓度真值，$y_R^i$表示右叶节点第i个DXN排放浓度真值，$c_L$表示左叶节点DXN排放浓度预测输出，$c_R$表示右叶节点DXN排放浓度预测输出；

通过最小化$\Omega_i(s,v)$，将训练集$(\boldsymbol{X}_{\text{Bootstrap}}^{n,j},\boldsymbol{y}_{\text{Bootstrap}}^{n,j})$切分为两个树节点，如下：

$$\min\left\{\Omega_i(s,v)\right\}_{i=1}^{N_{\text{Raw}}\times M}\xrightarrow{\text{树节点分裂}}\begin{cases}R_{\text{L}}^{N_{\text{L}}\times M}\\R_{\text{R}}^{N_{\text{R}}\times M}\end{cases}\tag{4}$$

其中，$R_{\text{L}}^{N_{\text{L}}\times M}$和$R_{\text{R}}^{N_{\text{R}}\times M}$表示切分后左右两个树节点所包含的样本集，$N_L$和$N_R$分别表示$R_{\text{L}}^{N_{\text{L}}\times M}$和$R_{\text{R}}^{N_{\text{R}}\times M}$中的样本数量；

当前左右树节点的DXN排放浓度预测输出值输出值$c_L^{\text{RF}}$和$c_R^{\text{RF}}$为样本真值的期望，如下：

$$\begin{cases}c_{\text{L}}^{\text{RF}}=E[\boldsymbol{y}_{\text{L}}],\ \ \boldsymbol{y}_{\text{L}}\in R_{\text{L}}^{N_{\text{L}}\times M}\\c_{\text{R}}^{\text{RF}}=E[\boldsymbol{y}_{\text{R}}],\ \ \boldsymbol{y}_{\text{R}}\in R_{\text{R}}^{N_{\text{R}}\times M}\end{cases}\tag{5}$$

其中，$y_L$和$y_R$表示$R_{\text{L}}^{N_{\text{L}}\times M}$和$R_{\text{R}}^{N_{\text{R}}\times M}$中的DXN排放浓度真值向量，$E[y_L]$和$E[y_R]$表示$y_L$和$y_R$的数学期望；

与RF不同，CRF中决策树分裂采用完全随机选择方式，表示为，

$$rand\left\{(s,v)_i\right\}_{i=1}^{N_{\text{Raw}}\times M}\xrightarrow{\text{树节点分裂}}\begin{cases}R_{\text{L}}^{N_{\text{L}}\times M}\\R_{\text{R}}^{N_{\text{R}}\times M}\end{cases}\tag{6}$$

其中，$rand\left\{(s,v)_i\right\}_{i=1}^{N_{\text{Raw}}\times M}$表示完全随机选取第sth个特征的值v作为切分点；

被随机分裂的左右树节点的DXN排放浓度预测输出值$c_{\text{L}}^{\text{CRF}}$和$c_{\text{R}}^{\text{CRF}}$为样本真值的期望，如下：

$$\begin{aligned}c_{\text{L}}^{\text{CRF}}&=E[\boldsymbol{y}_{\text{L}}],\ \ \boldsymbol{y}_{\text{L}}\in R_{\text{L}}^{N_{\text{L}}\times M}\\c_{\text{R}}^{\text{CRF}}&=E[\boldsymbol{y}_{\text{R}}],\ \ \boldsymbol{y}_{\text{R}}\in R_{\text{R}}^{N_{\text{R}}\times M}\end{aligned}\tag{7}$$

通过上述过程，第nth个混合森林组$f_n^{\text{FML}}(\cdot)$可表示为，

$$f_n^{\text{FML}}(\cdot)=\left\{f_{n,\text{RF}}^{\text{FML}}(\cdot),f_{n,\text{CRF}}^{\text{FML}}(\cdot)\right\}\tag{8}$$

其中，$f_{n,\text{RF}}^{\text{FML}}(\cdot)$表示第nth个随机森林，$f_{n,\text{CRF}}^{\text{FML}}(\cdot)$表示第nth个完全随机森林；

进而，第nth个映射特征$Z_n$可表示为

$$\begin{aligned}\boldsymbol{Z}_n&=f_n^{\text{FML}}(\boldsymbol{X})=\left\{f_{n,\text{RF}}^{\text{FML}}(\boldsymbol{X}),f_{n,\text{CRF}}^{\text{FML}}(\boldsymbol{X})\right\}\\&=[(c_{1,l}^{n,\text{RF}},c_{1,l}^{n,\text{RF}}),\ldots,(c_{n_{\text{Raw}},l}^{n,\text{RF}},c_{n_{\text{Raw}},l}^{n,\text{RF}}),\ldots,(c_{N_{\text{Raw}},l}^{n,\text{RF}},c_{N_{\text{Raw}},l}^{n,\text{RF}})]\end{aligned}\tag{9}$$

其中，$(c_{1,l}^{n,\text{RF}},c_{1,l}^{n,\text{RF}})$表示第nth组混合森林对来源于MSWI过程六个不同阶段的原始输入数据第1个样本的映射特征，$(c_{n_{\text{Raw}},l}^{n,\text{RF}},c_{n_{\text{Raw}},l}^{n,\text{RF}})$表示第nth组混合森林对来源于MSWI过程六个不同阶段的原始输入数据第$n_{\text{Raw}}$th个样本的映射特征，$(c_{N_{\text{Raw}},l}^{n,\text{RF}},c_{N_{\text{Raw}},l}^{n,\text{RF}})$表示第nth组混合森林对来

源于MSWI过程六个不同阶段的原始输入数据第$N_{Raw}$th个样本的映射特征；

最终，特征映射层的输出表示为：

$$\boldsymbol{Z}^{N}=\left(\boldsymbol{Z}_{1},\boldsymbol{Z}_{2},...,\boldsymbol{Z}_{N}\right)\in R^{N_{\mathrm{Raw}}\times 2N} \tag{10}$$

其中，$Z_1$为第1个映射特征，$Z_2$为第2个映射特征，$Z_N$为第N个映射特征，映射特征矩阵$Z^N$包含$N_{Raw}$个样本和2N维特征。

3.根据权利要求2所述的基于宽度混合森林回归的MSWI过程二噁英排放软测量方法，其特征在于，所述步骤S2，构建潜在特征提取层，依据贡献率对全联接混合矩阵的特征空间进行潜在特征提取，基于信息度量准则保证潜在有价值信息的最大化传递和最小化冗余，降低模型复杂度和计算消耗，具体包括：

首先，来源于MSWI过程六个不同阶段的原始输入数据X与特征映射矩阵$Z^N$组合得到全联接混合矩阵A，表示为：

$$\boldsymbol{A}=[\boldsymbol{X}\mid\boldsymbol{Z}^{N}]\in R^{N_{\mathrm{Raw}}\times(M+2N)} \tag{11}$$

其中，A含$N_{Raw}$个样本和(M+2N)维特征；

接着，考虑到A的维数远高于原始数据，此处利用PCA最小化A中的冗余信息，计算A的相关矩阵R，如下：

$$\boldsymbol{R}=\frac{1}{N_{\mathrm{Raw}}-1}\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}\in R^{(M+2N)\times(M+2N)} \tag{12}$$

进一步，对R进行奇异值分解，得到(M+2N)个特征值和相应特征向量，如下：

$$R=U_{(M+2N)}\,\Sigma_{(M+2N)}\,V_{(M+2N)} \tag{13}$$

其中，$U_{(M+2N)}$表示(M+2N)阶正交矩阵，$\Sigma_{(M+2N)}$表示(M+2N)阶对角矩阵，$V_{(M+2N)}$表示(M+2N)阶正交矩阵；

$$\boldsymbol{\Sigma}_{(M+2N)}=\begin{bmatrix}\sigma_{1} & & \\ & \ddots & \\ & & \sigma_{(M+2N)}\end{bmatrix} \tag{14}$$

其中，$\sigma_1>\sigma_2>\cdots>\sigma_{(M+2N)}$表示由大到小排列的特征值；

然后，根据设定潜在特征贡献阈值η，确定最终的主成分数量，

$$\eta=\sum_{q=1}^{Q_{\mathrm{PCA}}}\sigma_{q}\Big/\sum_{q=1}^{(M+2N)}\sigma_{q} \tag{15}$$

其中，潜在特征数量$Q_{PCA}<<(M+2N)$；

基于上述确定的$Q_{PCA}$个潜在特征，获得特征值集合$\left\{\sigma_q\right\}_{q=1}^{Q_{\mathrm{PCA}}}$对应的特征向量矩阵$\boldsymbol{V}_{Q_{\mathrm{PCA}}}$，即A的投影矩阵；然后，对A进行特征投影以实现冗余信息的最小化处理，将获得潜在特征记为$X^{PCA}$，即

$$\boldsymbol{X}^{\mathrm{PCA}}=\boldsymbol{A}\boldsymbol{V}_{I_{\mathrm{PCA}}}\in R^{N_{\mathrm{Raw}}\times M_{\mathrm{PCA}}} \tag{16}$$

其中，$\boldsymbol{V}_{Q_{\mathrm{PCA}}}\in R^{(M+2N)\times Q_{\mathrm{PCA}}}$表示前$Q_{PCA}$个潜在特征的特征向量；

进一步，计算所选潜在特征$X^{PCA}$与真值$\boldsymbol{y}\in\boldsymbol{R}^{N_{\mathrm{Raw}}\times 1}$间的互信息值$I^{MI}$，如下：

$$I^{\mathrm{MI}}(\boldsymbol{X}^{\mathrm{PCA}},\boldsymbol{y})=\sum_{q=1}^{Q_{\mathrm{PCA}}}p(\boldsymbol{x}_q^{\mathrm{PCA}},\boldsymbol{y})\log_2\frac{p(\boldsymbol{x}_q^{\mathrm{PCA}},\boldsymbol{y})}{p(\boldsymbol{x}_q^{\mathrm{PCA}}),p(\boldsymbol{y})} \tag{17}$$

其中，$p(\boldsymbol{x}_q^{\mathrm{PCA}},\boldsymbol{y})$表示第qth个潜在特征$\boldsymbol{x}_q^{\mathrm{PCA}}$与DXN排放浓度真值y的联合概率分布，$p(\boldsymbol{x}_q^{\mathrm{PCA}})$表示第qth个潜在特征$\boldsymbol{x}_q^{\mathrm{PCA}}$的边缘概率分布，p(y)表示DXN排放浓度真值y的边缘概率分布；

接着，通过信息最大化选择机制以保证所选择潜在特征与真值的相关性，表示为：

$$\left\{I_q^{\mathrm{MI}}\right\}_{q=1}^{Q_{\mathrm{PCA}}}\xrightarrow{I_q^{\mathrm{MI}}\geq\zeta}\left\{I_q^{\mathrm{MI}}\right\}_{q=1}^{Q_{\mathrm{PCA}}^{\mathrm{MI}}} \tag{18}$$

其中，$\left\{I_q^{\mathrm{MI}}\right\}_{q=1}^{Q}$表示$Q_{\mathrm{PCA}}$个潜在特征$\boldsymbol{x}_q^{\mathrm{PCA}}$与真值y的互信息值，ζ表示最大化信息的阈值，$\left\{I_q^{\mathrm{MI}}\right\}_{q=1}^{Q_{\mathrm{PCA}}^{\mathrm{MI}}}$表示与DXN排放浓度真值y信息相关度最大的$Q_{\mathrm{PCA}}^{\mathrm{MI}}$个潜在特征；

最终，获得包括$Q_{\mathrm{PCA}}^{\mathrm{MI}}$个潜在特征的新数据集$\{\boldsymbol{X}',\boldsymbol{y}\}\in R^{N_{\mathrm{Raw}}\times\left(Q_{\mathrm{PCA}}^{\mathrm{MI}}+1\right)}$，并设定提取后维数$M_{\mathrm{PCA}}^{\mathrm{MI}}=Q_{\mathrm{PCA}}^{\mathrm{MI}}$。

4.根据权利要求3所述的基于宽度混合森林回归的MSWI过程二噁英排放软测量方法，其特征在于，所述步骤S3中，构建特征增强层，基于所提取的潜在特征训练特征增强层以进一步增强特征表征能力，具体包括：

首先对新数据集{X′,y}进行基于Bootstrap和RSM的采样，获取混合森林算法的第个J训练子集，如下：

$$\left\{\boldsymbol{X}'^{k,j}_{\mathrm{Bootstrap}},\boldsymbol{y}^{k,j}_{\mathrm{Bootstrap}}\right\}_{j=1}^{J}=\varphi_k^{\mathrm{FEL}}\left(\phi_k^{\mathrm{FEL}}\left(\{\boldsymbol{X}',\boldsymbol{y}\},P_{\mathrm{Bootstrap}}\right)\right) \tag{19}$$

其中，$\boldsymbol{X}'^{k,j}_{\mathrm{Bootstrap}}$和$\boldsymbol{y}^{k,j}_{\mathrm{Bootstrap}}$为第个J训练子集的输入和输出，X′和y为新训练集的输入和输出，$\phi_k^{\mathrm{FEL}}(\cdot)$表示对第kth个混合森林组的Bootstrap采样，$\varphi_k^{\mathrm{FEL}}(\cdot)$表示对第kth个混合森林组的RSM采样；

接着，以第kth个混合森林组中第j个RF的构建为例，如下：

$$\left\{\boldsymbol{X}'^{k,j}_{\mathrm{Bootstrap}},\boldsymbol{y}^{k,j}_{\mathrm{Bootstrap}}\right\}\xrightarrow{\Omega_j(s,v)}f_{k,j}^{\mathrm{DT\text{-}RF}}(\cdot)=\sum_{l=1}^{L}c_l I(\boldsymbol{x}^{k,j}_{\mathrm{Bootstrap}}\in R_l),\quad l=1,2,...,L \tag{20}$$

其中，$f_{k,j}^{\mathrm{DT\text{-}RF}}(\cdot)$表示特征增强层中第kth个混合森林组中RF的第jth个决策树；L表示决策树叶节点的数量；$c_1$采用递归分裂方式计算，具体过程公式(3)-(5)；

进而，可得到特征增强层中第kth个混合森林组中的RF模型，其表示为，

$$f_{k,\mathrm{RF}}^{\mathrm{FEL}}(\cdot)=\left\{f_{k,j}^{\mathrm{DT\text{-}RF}}(\cdot)\right\}_{j=1}^{J} \tag{21}$$

然后，类似地以第kth个混合森林组中的第j个CRF的构建为例，如下：

$$\left\{\boldsymbol{X}'^{k,j}_{\mathrm{Bootstrap}},\boldsymbol{y}^{k,j}_{\mathrm{Bootstrap}}\right\}\xrightarrow{rand_j(s,v)}f_{k,j}^{\mathrm{DT\text{-}CRF}}(\cdot)=\sum_{l=1}^{L}c_l I(\boldsymbol{x}^{k,j}_{\mathrm{Bootstrap}}\in R_l),\quad l=1,2,...,L \tag{22}$$

其中，$f_{k,j}^{\mathrm{DT\text{-}CRF}}(\cdot)$表示特征增强层中第kth个混合森林组中CRF的第jth个决策树；$c_1$采用递归分裂方式计算，具体过程见公式(6)-(7)；

进而，可得到特征增强层中第kth个混合森林组的CRF模型，其表示为，

$$f_{k,\mathrm{CRF}}^{\mathrm{FEL}}(\cdot)=\left\{f_{k,j}^{\mathrm{DT\text{-}CRF}}(\cdot)\right\}_{j=1}^{J} \tag{23}$$

通过上述过程,得到第kth个混合森林组$f_k^{\mathrm{FEL}}(\cdot)$;进而,第kth个增强特征可表示如下:

$$\begin{aligned}\boldsymbol{H}_k &= f_k^{\mathrm{FEL}}(\boldsymbol{X}')=\left[f_{k,\mathrm{RF}}^{\mathrm{FEL}}(\boldsymbol{X}'),f_{k,\mathrm{CRF}}^{\mathrm{FEL}}(\boldsymbol{X}')\right]\\ &=\left[(c_{1,l}^{k,\mathrm{RF}},c_{1,l}^{k,\mathrm{CRF}}),\ldots,(c_{n_{\mathrm{Raw}},l}^{k,\mathrm{RF}},c_{n_{\mathrm{Raw}},l}^{k,\mathrm{RF}}),\ldots,(c_{N_{\mathrm{Raw}},l}^{k,\mathrm{RF}},c_{N_{\mathrm{Raw}},l}^{k,\mathrm{CRF}})\right]\end{aligned} \tag{24}$$

其中,$(c_{1,l}^{k,\mathrm{RF}},c_{1,l}^{k,\mathrm{RF}})$表示第kth个混合森林组对新数据中第1个样本的增强映射,$(c_{n_{\mathrm{Raw}},l}^{k,\mathrm{RF}},c_{n_{\mathrm{Raw}},l}^{k,\mathrm{RF}})$表示第kth个混合森林组对新数据中第$n_{Raw}$th个样本的增强映射,$(c_{N_{\mathrm{Raw}},l}^{k,\mathrm{RF}},c_{N_{\mathrm{Raw}},l}^{k,\mathrm{RF}})$表示第kth个混合森林组对新数据中第$N_{Raw}$th个样本的增强映射;

最后,特征增强层的输出$H^K$表示如下:

$$\boldsymbol{H}^K=[\boldsymbol{H}_1,\boldsymbol{H}_2,\ldots,\boldsymbol{H}_K]\in R^{N_{\mathrm{Raw}}\times 2K} \tag{25}$$

其中,$H_1$为第1个增强特征,$H_2$为第2个增强特征,$H_K$为第K个增强特征;

当不考虑增量学习策略时,BHFR模型的表示如下:

$$\begin{aligned}\boldsymbol{Y}&=\boldsymbol{G}^K\boldsymbol{W}^K\\ &=[\boldsymbol{Z}_1,\boldsymbol{Z}_2,\ldots,\boldsymbol{Z}_N\,|\,\boldsymbol{H}_1,\boldsymbol{H}_2,\ldots\boldsymbol{H}_K]\boldsymbol{W}^K\end{aligned} \tag{26}$$

其中,$G^K$表示特征映射层与特征增强层输出的组合,即$G^K=[Z^N|H^K]$,其包含$N_{Raw}$个样本和(2N+2K)维特征;$W^K$表示特征映射层和特征增强层与输出层间的权重,其计算如下:

$$W^K=(\lambda I+[G^K]^T G^K)^{-1}[G^K]^T Y \tag{27}$$

其中,I表示单位矩阵,λ表示正则项系数;相应地,$G^K$的伪逆计算可表示为:

$$\begin{aligned}[\boldsymbol{G}^K]^{\dagger}&=(\lambda\boldsymbol{I}+[\boldsymbol{G}^K]^{\mathrm{T}}\boldsymbol{G})^{-1}[\boldsymbol{G}^K]^{\mathrm{T}}\\ &=[\boldsymbol{Z}^N\,|\,\boldsymbol{H}^K]^{\dagger}\end{aligned} \tag{28}$$

5.根据权利要求4所述的基于宽度混合森林回归的MSWI过程二噁英排放软测量方法,其特征在于,所述步骤S4,构建增量学习层,通过增量式学习策略构建增量学习层,采用Moore-Penrose伪逆获得权重矩阵,进而实现BHFR软测量模型的高精度建模,具体包括:

首先,对新数据集{X′,y}进行基于Bootstrap和RSM的采样,获取混合森林算法训练子集,过程如下:

$$\left\{\boldsymbol{X}'^{p,j}_{\mathrm{Bootstrap}},\boldsymbol{y}^{p,j}_{\mathrm{Bootstrap}}\right\}_{j=1}^{J}=\varphi_p^{\mathrm{ILL}}\left(\phi_p^{\mathrm{ILL}}\left\{\{\boldsymbol{X}',\boldsymbol{y}\},P_{\mathrm{Bootstrap}}\right\}\right) \tag{29}$$

其中,$\boldsymbol{X}'^{p,j}_{\mathrm{Bootstrap}}$和$\boldsymbol{y}^{p,j}_{\mathrm{Bootstrap}}$为混合森林算法第个J训练子集的输入和输出,X′和y为新训练集的输入和输出,$\phi_i^{\mathrm{ILL}}(\cdot)$和$\varphi_p^{\mathrm{ILL}}(\cdot)$表示增量学习层中第pth个混合森林组的Bootstrap采样和RSM采样;

接着,构建第pth个混合森林组中的决策树$f_{p,\mathrm{RF}}^{\mathrm{ILL}}(\cdot)$和$f_{p,\mathrm{CRF}}^{\mathrm{ILL}}(\cdot)$,其过程与特征映射层和特征增量层相同,此处不再赘述;

进一步,当增加1个混合森林组后,特征映射层、特征增量层和增量学习层的输出$G^{K+1}$表示如下:

$$\begin{aligned}\boldsymbol{G}^{K+1}&=\left[\boldsymbol{G}^{K}\,|\,f_{1}^{\mathrm{FEL}}(\boldsymbol{X}')\right]\\&=\left[\boldsymbol{G}^{K}\,|\left\{f_{1,\mathrm{RF}}^{\mathrm{FEL}}(\boldsymbol{X}'),f_{1,\mathrm{CRF}}^{\mathrm{FEL}}(\boldsymbol{X}')\right\}\right]\\&=\left[\boldsymbol{G}^{K}\,|[(c_{1,l}^{1,\mathrm{RF}},c_{1,l}^{1,\mathrm{CRF}}),\ldots,(c_{N_{\mathrm{Raw}},l}^{1,\mathrm{RF}},c_{N_{\mathrm{Raw}},l}^{1,\mathrm{CRF}})]\right]\end{aligned}\tag{30}$$

其中，$G^{k}=[Z^{n}|H^{k}]$包含$N_{Raw}$个样本和(2N+2K)维特征，$G^{K+1}$包含$N_{Raw}$个样本和(2N+2K+2J)维特征；

然后，进行$G^{K+1}$的Moore-Penrose逆矩阵的递推更新，如下：

$$\boldsymbol{B}^{\mathrm{T}}=\begin{cases}[\boldsymbol{C}]^{\dagger}, & \text{if } \boldsymbol{C}\neq 0\\ [1+\boldsymbol{D}^{\mathrm{T}}\boldsymbol{D}]^{-1}\boldsymbol{D}^{\mathrm{T}}[\boldsymbol{G}^{k}]^{\dagger}, & \text{if } \boldsymbol{C}=0\end{cases}\tag{31}$$

其中，矩阵C和矩阵D的计算如下：

$$C=H_{K+1}-G^{K}D\tag{32}$$

$$\boldsymbol{D}=[\boldsymbol{G}^{K}]^{\dagger}f_{1}^{\mathrm{ILL}}(\boldsymbol{X}^{\mathrm{New}})\tag{33}$$

进而，$G^{K+1}$的Moore-Penrose逆矩阵的递推公式如下：

$$[\boldsymbol{G}^{K+1}]^{\dagger}=\begin{bmatrix}[\boldsymbol{G}^{K}]^{\dagger}-\boldsymbol{D}\boldsymbol{B}^{\mathrm{T}}\\ \boldsymbol{B}^{\mathrm{T}}\end{bmatrix}\tag{34}$$

进一步，计算特征映射层、特征增量层和增量学习层与输出层间权重的更新矩阵$W^{K+1}$，如下：

$$\boldsymbol{W}^{K+1}=\begin{bmatrix}\boldsymbol{W}^{K}-\boldsymbol{D}\boldsymbol{B}^{\mathrm{T}}\boldsymbol{Y}\\ \boldsymbol{B}^{\mathrm{T}}\boldsymbol{Y}\end{bmatrix}\tag{35}$$

其中，$W^{K}=(\lambda I+[G^{K}]^{T}G^{K})^{-1}[G^{K}]^{T}Y$；

由于采用上述伪逆更新策略只需要计算增量学习层混合森林组的伪逆矩阵，因此能够实现快速的增量式学习；

进一步，根据训练误差的收敛程度实现自适应增量学习；

定义误差的收敛阈值为$\theta_{Con}$用以确定增量学习中混合森林组的数量p；相应地，BHFR模型的增量学习训练误差表示如下：

$$\begin{aligned}&\ell=\lim_{p\to\infty}\left|\frac{1}{N}\left(\sqrt{\left(\boldsymbol{G}^{K+p}\boldsymbol{W}^{K+p}-\boldsymbol{y}\right)^{2}}-\sqrt{\left(\boldsymbol{G}^{K+p+1}\boldsymbol{W}^{K+p+1}-\boldsymbol{y}\right)^{2}}\right)\right|\leq\theta_{\mathrm{Con}}\\&\text{S.T. }\ \theta_{\mathrm{Con}}\geq 0\end{aligned}\tag{36}$$

其中，1表示增量学习第p+1个与第p个混合森林组的训练误差值，$\sqrt{\left(\boldsymbol{G}^{K+p}\boldsymbol{W}^{K+p}-\boldsymbol{y}\right)^{2}}$和$\sqrt{\left(\boldsymbol{G}^{K+p+1}\boldsymbol{W}^{K+p+1}-\boldsymbol{y}\right)^{2}}$表示包含p个和p+1个混合森林组的BHFR模型训练误差；

最终，所提BHFR软测量模型的预测输出$\hat{\boldsymbol{Y}}$为：

$$\hat{\boldsymbol{Y}}=\boldsymbol{G}^{K+P}\boldsymbol{W}^{K+P}\tag{37}$$

# 基于宽度混合森林回归的MSWI过程二噁英排放软测量方法

## 技术领域

[0001] 本发明涉及二噁英排放软测量技术领域,特别是涉及一种基于宽度混合森林回归的MSWI过程二噁英排放软测量方法。

## 背景技术

[0002] 城市固废焚烧(Municipal SolidWaste Incineration,MSWI)是目前世界范围内解决城市“垃圾围城”困境的主要方式之一,具有无害化、减量化和资源化等显著优势。二噁英(Dioxin,DXN)作为MSWI过程排放的有组织废气中具有持久性和剧毒性的有机污染物,是造成焚烧建厂存在“邻避现象”的主要原因,也是MSWI过程必须最小化控制的重要环保指标之一。基于高分辨气相色谱-高分辨质谱(HRGC/HRMS)的离线化验分析方法是目前用于检测DXN排放浓度的主要手段,存在技术难度大、时间滞后性大、人力与经济成本高等缺点,已经成为阻碍MSWI过程实现实时优化控制的关键因素之一。因此,DXN排放浓度的在线检测已成为MSWI过程的首要挑战问题。

[0003] 针对上述问题,利用可在线检测的DXN关联物构建关联模型进而间接获得DXN浓度的在线间接检测方法成为热点;然而,其存在设备复杂、成本高、干扰因素多、预测精度无法保证等问题,同时其在本质上也是一种结合数据建模的检测手段。相较于离线分析和在线间接检测方法而言,基于工业集散控制系统采集的易检测过程数据驱动的软测量技术是解决DXN无法在线检测问题的有效途径,具有稳定、精准和快速响应等特点。软测量技术已在石油、化工和炼钢等复杂工业过程的难测参数检测中广泛应用。

## 发明内容

[0004] 本发明的目的是提供一种基于宽度混合森林回归的MSWI过程二噁英排放软测量方法,以MSWI过程DXN排放浓度检测为目标,提出了基于宽度混合森林回归(Broad HybridForest Regression,BHFR)的软测量建模算法。

[0005] 为实现上述目的,本发明提供了如下方案:

[0006] 一种基于宽度混合森林回归的MSWI过程二噁英排放软测量方法,基于BLS框架,以非微分基学习器替换神经元构建面向小样本高维数据的BHFR软测量模型,所述BHFR软测量模型包括特征映射层、潜在特征提取层、特征增强层和增量学习层的构建,具体包括以下步骤:

[0007] S1,构建特征映射层,构建由随机森林RF和完全随机森林CRF组成的混合森林组对高维特征进行映射;

[0008] S2,构建潜在特征提取层,依据贡献率对全联接混合矩阵的特征空间进行潜在特征提取,基于信息度量准则保证潜在有价值信息的最大化传递和最小化冗余,降低模型复杂度和计算消耗;

[0009] S3,构建特征增强层,基于所提取的潜在特征训练特征增强层以进一步增强特征表征能力;

[0010] S4,构建增量学习层,通过增量式学习策略构建增量学习层,采用Moore-Penrose伪逆获得权重矩阵,进而实现BHFR软测量模型的高精度建模;

[0011] S5,采用高维基准数据集和工业过程DXN数据集验证所述软测量模型;

[0012] S6,采用步骤S1-S5建立的软测量模型,对MSWI过程二噁英排放进行软测量。

[0013] 进一步的,所述步骤S1,构建特征映射层,构建由随机森林RF和完全随机森林CRF组成的混合森林组对高维特征进行映射,具体包括:

[0014] 设原始数据为{X,y},其中 $\boldsymbol{X} \in R^{N_{\text{Raw}} \times M}$ 是原始输入数据,$N_{Raw}$是原始数据的数量,M是原始输入数据的维数,其来源于MSWI过程的六个不同阶段,以秒为单位在DCS系统采集与存储,$\boldsymbol{y} \in R^{N_{\text{Raw}} \times 1}$ 是DXN排放浓度的输出真值,其来源于采用离线检测法得到排放物DXN检测样本;以特征映射层的第nth个混合森林组为例描述特征映射层的建模过程:

[0015] 对{X,y}进行Bootstrap和随机子空间RSM采样,获得混合森林组模型的J个训练子集,如下:

[0016] $$\left\{\boldsymbol{X}_{\text{Bootstrap}}^{n,j}, \boldsymbol{y}_{\text{Bootstrap}}^{n,j}\right\}_{j=1}^{J} = \varphi_n^{\text{FML}}\left(\phi_n^{\text{FML}}((\boldsymbol{X},\boldsymbol{y}), P_{\text{Bootstrap}})\right) \tag{1}$$

[0017] 其中,$\boldsymbol{X}_{\text{Bootstrap}}^{n,j}$ 和 $\boldsymbol{y}_{\text{Bootstrap}}^{n,j}$ 为第J个训练子集的输入和输出,$\phi_n^{\text{FML}}(\cdot)$ 和 $\varphi_n^{\text{FML}}(\cdot)$ 表示特征映射层中对第nth个混合森林组的Bootstrap和RSM采样,$P_{Bootstrap}$表示Bootstrap采样概率;

[0018] 基于 $\left\{\boldsymbol{X}_{\text{Bootstrap}}^{n,j}, \boldsymbol{y}_{\text{Bootstrap}}^{n,j}\right\}_{j=1}^{J}$ 训练包含J个决策树的混合森林算法,其中特征映射层中的第nth个混合森林组的第jth个决策树表示如下:

[0019] $$f_{n,j}^{\text{DT}}(\cdot) = \sum_{l=1}^{L} c_l I(\boldsymbol{x}_{\text{Bootstrap}}^{n,j} \in R_l), \quad l = 1,2,...,L \tag{2}$$

[0020] 其中,L表示决策树叶节点数量,I(•)表示指示函数,$c_l$采用递归分裂方式计算;

[0021] RF中决策树的分裂损失函数$\Omega_i(\cdot)$表示为:

[0022] $$\begin{aligned}\Omega_i(s,v) &= \min\left([\boldsymbol{y}_L - E[\boldsymbol{y}_L]] + [\boldsymbol{y}_R - E[\boldsymbol{y}_R]]\right) \\ &= \min\left(\sum_{x_{\text{Bootstrap}}^{n,j} \in R_{\text{L}}} (y_{\text{L}}^{\text{i}} - c_{\text{L}})^2 + \sum_{x_{\text{Bootstrap}}^{n,j} \in R_{\text{R}}} (y_{\text{R}}^{\text{i}} - c_{\text{R}})^2\right)\end{aligned} \tag{3}$$

[0023] 其中,$\Omega_i(s,v)$表示第sth个特征的值v作为切分准则的损失函数值,$y_L$表示左叶节点的DXN排放浓度真值向量,E[$y_L$]表示$y_L$的数学期望,$y_R$表示右叶节点的DXN排放浓度真值向量,E[$y_R$]表示$y_R$的数学期望,$y_{\text{L}}^{\text{i}}$ 表示左叶节点第i个DXN排放浓度真值,$y_{\text{R}}^{\text{i}}$ 表示右叶节点第i个DXN排放浓度真值,$c_L$表示左叶节点DXN排放浓度预测输出,$c_R$表示右叶节点DXN排放浓度预测输出;

[0024] 通过最小化$\Omega_i(s,v)$,将训练集$(\boldsymbol{X}_{\text{Bootstrap}}^{n,j}, \boldsymbol{y}_{\text{Bootstrap}}^{n,j})$切分为两个树节点,如下:

[0025] $$\min\left\{\Omega_i(s,v)\right\}_{i=1}^{N_{\text{Raw}} \times M} \xrightarrow{\text{树节点分裂}} \begin{cases} R_{\text{L}}^{N_{\text{L}} \times M} \\ R_{\text{R}}^{N_{\text{R}} \times M} \end{cases} \tag{4}$$

[0026] 其中,$R_{\text{L}}^{N_{\text{L}} \times M}$ 和 $R_{\text{R}}^{N_{\text{R}} \times M}$ 表示切分后左右两个树节点所包含的样本集,$N_L$和$N_R$分别表示 $R_{\text{L}}^{N_{\text{L}} \times M}$ 和 $R_{\text{R}}^{N_{\text{R}} \times M}$ 中的样本数量;

[0027] 当前左右树节点的DXN排放浓度预测输出值输出值$c_{\mathrm{L}}^{\mathrm{RF}}$和$c_{\mathrm{R}}^{\mathrm{RF}}$为样本真值的期望，如下：

[0028] $$\begin{cases} c_{\mathrm{L}}^{\mathrm{RF}} = E[\boldsymbol{y}_{\mathrm{L}}], \ \boldsymbol{y}_{\mathrm{L}} \in R_{\mathrm{L}}^{N_{\mathrm{L}} \times M} \\ c_{\mathrm{R}}^{\mathrm{RF}} = E[\boldsymbol{y}_{\mathrm{R}}], \ \boldsymbol{y}_{\mathrm{R}} \in R_{\mathrm{R}}^{N_{\mathrm{R}} \times M} \end{cases} \tag{5}$$

[0029] 其中，$y_L$和$y_R$表示$\mathbb{R}_{\mathrm{L}}^{N_{\mathrm{L}} \times M}$和$\mathbb{R}_{\mathrm{R}}^{N_{\mathrm{R}} \times M}$中的DXN排放浓度真值向量，E[$y_L$]和E[$y_R$]表示$y_L$和$y_R$的数学期望；

[0030] 与RF不同，CRF中决策树分裂采用完全随机选择方式，表示为，

[0031] $$rand\left\{(s,v)_i\right\}_{i=1}^{N_{\mathrm{Raw}} \times M} \xrightarrow{\text{树节点分裂}} \begin{cases} R_{\mathrm{L}}^{N_{\mathrm{L}} \times M} \\ R_{\mathrm{R}}^{N_{\mathrm{R}} \times M} \end{cases} \tag{6}$$

[0032] 其中，$rand\left\{(s,v)_i\right\}_{i=1}^{N_{\mathrm{Raw}} \times M}$表示完全随机选取第sth个特征的值v作为切分点；

[0033] 被随机分裂的左右树节点的DXN排放浓度预测输出值$c_{\mathrm{L}}^{\mathrm{CRF}}$和$c_{\mathrm{R}}^{\mathrm{CRF}}$为样本真值的期望，如下：

[0034] $$\begin{aligned} c_{\mathrm{L}}^{\mathrm{CRF}} &= E[\boldsymbol{y}_{\mathrm{L}}], \ \boldsymbol{y}_{\mathrm{L}} \in R_{\mathrm{L}}^{N_{\mathrm{L}} \times M} \\ c_{\mathrm{R}}^{\mathrm{CRF}} &= E[\boldsymbol{y}_{\mathrm{R}}], \ \boldsymbol{y}_{\mathrm{R}} \in R_{\mathrm{R}}^{N_{\mathrm{R}} \times M} \end{aligned} \tag{7}$$

[0035] 通过上述过程，第nth个混合森林组$f_n^{\mathrm{FML}}(\cdot)$可表示为，

[0036] $$f_n^{\mathrm{FML}}(\cdot) = \left\{ f_{n,\mathrm{RF}}^{\mathrm{FML}}(\cdot), f_{n,\mathrm{CRF}}^{\mathrm{FML}}(\cdot) \right\} \tag{8}$$

[0037] 其中，$f_{n,\mathrm{RF}}^{\mathrm{FML}}(\cdot)$表示第nth个随机森林，$f_{n,\mathrm{CRF}}^{\mathrm{FML}}(\cdot)$表示第nth个完全随机森林；进而，第nth个映射特征$Z_n$可表示为

[0038] $$\begin{aligned} \boldsymbol{Z}_n &= f_n^{\mathrm{FML}}(\boldsymbol{X}) = \left\{ f_{n,\mathrm{RF}}^{\mathrm{FML}}(\boldsymbol{X}), f_{n,\mathrm{CRF}}^{\mathrm{FML}}(\boldsymbol{X}) \right\} \\ &= [(c_{1,l}^{n,\mathrm{RF}}, c_{1,l}^{n,\mathrm{RF}}), \ldots, (c_{n_{\mathrm{Raw}},l}^{n,\mathrm{RF}}, c_{n_{\mathrm{Raw}},l}^{n,\mathrm{RF}}), \ldots, (c_{N_{\mathrm{Raw}},l}^{n,\mathrm{RF}}, c_{N_{\mathrm{Raw}},l}^{n,\mathrm{RF}})] \end{aligned} \tag{9}$$

[0039] 其中，$(c_{1,l}^{n,\mathrm{RF}}, c_{1,l}^{n,\mathrm{RF}})$表示第nth组混合森林对来源于MSWI过程六个不同阶段的原始输入数据第1个样本的映射特征，$(c_{n_{\mathrm{Raw}},l}^{n,\mathrm{RF}}, c_{n_{\mathrm{Raw}},l}^{n,\mathrm{RF}})$表示第nth组混合森林对来源于MSWI过程六个不同阶段的原始输入数据第$n_{Raw}$th个样本的映射特征，$(c_{N_{\mathrm{Raw}},l}^{n,\mathrm{RF}}, c_{N_{\mathrm{Raw}},l}^{n,\mathrm{RF}})$表示第nth组混合森林对来源于MSWI过程六个不同阶段的原始输入数据第$N_{Raw}$th个样本的映射特征；

[0040] 最终，特征映射层的输出表示为：

[0041] $$\boldsymbol{Z}^N = \left(\boldsymbol{Z}_1, \boldsymbol{Z}_2, ..., \boldsymbol{Z}_N\right) \in R^{N_{\mathrm{Raw}} \times 2N} \tag{10}$$

[0042] 其中，$Z_1$为第1个映射特征，$Z_2$为第2个映射特征，$Z_N$为第N个映射特征，映射特征矩阵$Z^N$包含$N_{Raw}$个样本和2N维特征。

[0043] 进一步的，所述步骤S2，构建潜在特征提取层，依据贡献率对全联接混合矩阵的特征空间进行潜在特征提取，基于信息度量准则保证潜在有价值信息的最大化传递和最小化冗余，降低模型复杂度和计算消耗，具体包括：

[0044] 首先，来源于MSWI过程六个不同阶段的原始输入数据X与特征映射矩阵$Z^N$组合得到全联接混合矩阵A，表示为：

[0045] $$\boldsymbol{A}=[\boldsymbol{X}\mid\boldsymbol{Z}^{N}]\in R^{N_{\mathrm{Raw}}\times(M+2N)} \tag{11}$$

[0046] 其中，A含$N_{Raw}$个样本和(M+2N)维特征；

[0047] 接着，考虑到A的维数远高于原始数据，此处利用PCA最小化A中的冗余信息，计算A的相关矩阵R，如下：

[0048] $$\boldsymbol{R}=\frac{1}{N_{\mathrm{Raw}}-1}\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}\in R^{(M+2N)\times(M+2N)} \tag{12}$$

[0049] 进一步，对R进行奇异值分解，得到(M+2N)个特征值和相应特征向量，如下：

[0050] $R=U_{(M+2N)}\Sigma_{(M+2N)}V_{(M+2N)}$ (13)

[0051] 其中，$U_{(M+2N)}$表示(M+2N)阶正交矩阵，$\Sigma_{(M+2N)}$表示(M+2N)阶对角矩阵，$V_{(M+2N)}$表示(M+2N)阶正交矩阵；

[0052] $$\boldsymbol{\Sigma}_{(M+2N)}=\begin{bmatrix}\sigma_1 & & \\ & \ddots & \\ & & \sigma_{(M+2N)}\end{bmatrix} \tag{14}$$

[0053] 其中，$\sigma_1>\sigma_2>\cdots>\sigma_{(M+2N)}$表示由大到小排列的特征值；

[0054] 然后，根据设定潜在特征贡献阈值η，确定最终的主成分数量，

[0055] $$\eta=\sum_{q=1}^{Q_{\mathrm{PCA}}}\sigma_q\Big/\sum_{q=1}^{(M+2N)}\sigma_q \tag{15}$$

[0056] 其中，潜在特征数量$Q_{PCA}<<(M+2N)$；

[0057] 基于上述确定的$Q_{PCA}$个潜在特征，获得特征值集合$\left\{\sigma_q\right\}_{q=1}^{Q_{\mathrm{PCA}}}$对应的特征向量矩阵$V_{QPCA}$，即A的投影矩阵；然后，对A进行特征投影以实现冗余信息的最小化处理，将获得潜在特征记为$X^{PCA}$，即

[0058] $$\boldsymbol{X}^{\mathrm{PCA}}=\boldsymbol{A}\boldsymbol{V}_{I_{\mathrm{PCA}}}\in R^{N_{\mathrm{Raw}}\times M_{\mathrm{PCA}}} \tag{16}$$

[0059] 其中，$\boldsymbol{V}_{Q_{\mathrm{PCA}}}\in R^{(M+2N)\times Q_{\mathrm{PCA}}}$表示前$Q_{PCA}$个潜在特征的特征向量；

[0060] 进一步，计算所选潜在特征$X^{PCA}$与真值$\boldsymbol{y}\in\boldsymbol{R}^{N_{\mathrm{Raw}}\times 1}$间的互信息值$I^{MI}$，如下：

[0061] $$I^{\mathrm{MI}}(\boldsymbol{X}^{\mathrm{PCA}},\boldsymbol{y})=\sum_{q=1}^{Q_{\mathrm{PCA}}}p(\boldsymbol{x}_q^{\mathrm{PCA}},\boldsymbol{y})\log_2\frac{p(\boldsymbol{x}_q^{\mathrm{PCA}},\boldsymbol{y})}{p(\boldsymbol{x}_q^{\mathrm{PCA}}),p(\boldsymbol{y})} \tag{17}$$

[0062] 其中，$p(\boldsymbol{x}_q^{\mathrm{PCA}},\boldsymbol{y})$表示第qth个潜在特征$\boldsymbol{x}_q^{\mathrm{PCA}}$与DXN排放浓度真值y的联合概率分布，$p(\boldsymbol{x}_q^{\mathrm{PCA}})$表示第qth个潜在特征$\boldsymbol{x}_q^{\mathrm{PCA}}$的边缘概率分布，p(y)表示DXN排放浓度真值y的边缘概率分布；

[0063] 接着，通过信息最大化选择机制以保证所选择潜在特征与真值的相关性，表示为：

[0064] $$\left\{I_q^{\mathrm{MI}}\right\}_{q=1}^{Q_{\mathrm{PCA}}}\xrightarrow{I_q^{\mathrm{MI}}\geq\zeta}\left\{I_q^{\mathrm{MI}}\right\}_{q=1}^{Q_{\mathrm{PCA}}^{\mathrm{MI}}} \tag{18}$$

[0065] 其中，$\left\{I_q^{\mathrm{MI}}\right\}_{q=1}^{Q}$表示$Q_{PCA}$个潜在特征$\boldsymbol{x}_q^{\mathrm{PCA}}$与真值y的互信息值，ζ表示最大化信息的阈值，$\left\{I_q^{\mathrm{MI}}\right\}_{q=1}^{Q_{\mathrm{PCA}}^{\mathrm{MI}}}$表示与DXN排放浓度真值y信息相关度最大的$Q_{\mathrm{PCA}}^{\mathrm{MI}}$个潜在特征；

[0066] 最终，获得包括$Q_{\mathrm{PCA}}^{\mathrm{MI}}$个潜在特征的新数据集$\{\boldsymbol{X}',\boldsymbol{y}\}\in R^{N_{\mathrm{Raw}}\times\left(Q_{\mathrm{PCA}}^{\mathrm{MI}}+1\right)}$，并设定提取后维数$M_{\mathrm{PCA}}^{\mathrm{MI}}=Q_{\mathrm{PCA}}^{\mathrm{MI}}$。

[0067] 进一步的，所述步骤S3中，构建特征增强层，基于所提取的潜在特征训练特征增强层以进一步增强特征表征能力，具体包括：

[0068] 首先对新数据集{X′,y}进行基于Bootstrap和RSM的采样，获取混合森林算法的第个J训练子集，如下：

[0069] $$\left\{\boldsymbol{X}'^{k,j}_{\mathrm{Bootstrap}},\boldsymbol{y}^{k,j}_{\mathrm{Bootstrap}}\right\}_{j=1}^{J}=\varphi_k^{\mathrm{FEL}}\left(\phi_k^{\mathrm{FEL}}\left(\{\boldsymbol{X}',\boldsymbol{y}\},P_{\mathrm{Bootstrap}}\right)\right)\tag{19}$$

[0070] 其中，$\boldsymbol{X}'^{k,j}_{\mathrm{Bootstrap}}$和$\boldsymbol{y}^{k,j}_{\mathrm{Bootstrap}}$为第个J训练子集的输入和输出，X′和y为新训练集的输入和输出，$\phi_k^{\mathrm{FEL}}(\cdot)$表示对第kth个混合森林组的Bootstrap采样，$\varphi_k^{\mathrm{FEL}}(\cdot)$表示对第kth个混合森林组的RSM采样；

[0071] 接着，以第kth个混合森林组中第j个RF的构建为例，如下：

[0072] $$\left\{\boldsymbol{X}'^{k,j}_{\mathrm{Bootstrap}},\boldsymbol{y}^{k,j}_{\mathrm{Bootstrap}}\right\}\xrightarrow{\Omega_j(s,v)}f_{k,j}^{\text{DT-RF}}(\cdot)=\sum_{l=1}^{L}c_l I(\boldsymbol{x}^{k,j}_{\mathrm{Bootstrap}}\in R_l),\quad l=1,2,...,L\tag{20}$$

[0073] 其中，$f_{k,j}^{\text{DT-RF}}(\cdot)$表示特征增强层中第kth个混合森林组中RF的第jth个决策树；L表示决策树叶节点的数量；$c_1$采用递归分裂方式计算，具体过程公式(3)-(5)；

[0074] 进而，可得到特征增强层中第kth个混合森林组中的RF模型，其表示为，

[0075] $$f_{k,\mathrm{RF}}^{\mathrm{FEL}}(\cdot)=\left\{f_{k,j}^{\text{DT-RF}}(\cdot)\right\}_{j=1}^{J}\tag{21}$$

[0076] 然后，类似地以第kth个混合森林组中的第j个CRF的构建为例，如下：

[0077] $$\left\{\boldsymbol{X}'^{k,j}_{\mathrm{Bootstrap}},\boldsymbol{y}^{k,j}_{\mathrm{Bootstrap}}\right\}\xrightarrow{rand_j(s,v)}f_{k,j}^{\text{DT-CRF}}(\cdot)=\sum_{l=1}^{L}c_l I(\boldsymbol{x}^{k,j}_{\mathrm{Bootstrap}}\in R_l),\quad l=1,2,...,L\tag{22}$$

[0078] 其中，$f_{k,j}^{\text{DT-CRF}}(\cdot)$表示特征增强层中第kth个混合森林组中CRF的第jth个决策树；$c_1$采用递归分裂方式计算，具体过程见公式(6)-(7)；

[0079] 进而，可得到特征增强层中第kth个混合森林组的CRF模型，其表示为，

[0080] $$f_{k,\mathrm{CRF}}^{\mathrm{FEL}}(\cdot)=\left\{f_{k,j}^{\text{DT-CRF}}(\cdot)\right\}_{j=1}^{J}\tag{23}$$

[0081] 通过上述过程，得到第kth个混合森林组$f_k^{\mathrm{FEL}}(\cdot)$；进而，第kth个增强特征可表示如下：

[0082] $$\begin{aligned}\boldsymbol{H}_k&=f_k^{\mathrm{FEL}}(\boldsymbol{X}')=\left[f_{k,\mathrm{RF}}^{\mathrm{FEL}}(\boldsymbol{X}'),f_{k,\mathrm{CRF}}^{\mathrm{FEL}}(\boldsymbol{X}')\right]\\&=\left[(c_{1,l}^{k,\mathrm{RF}},c_{1,l}^{k,\mathrm{CRF}}),\ldots,(c_{n_{\mathrm{Raw}},l}^{k,\mathrm{RF}},c_{n_{\mathrm{Raw}},l}^{k,\mathrm{RF}}),\ldots,(c_{N_{\mathrm{Raw}},l}^{k,\mathrm{RF}},c_{N_{\mathrm{Raw}},l}^{k,\mathrm{CRF}})\right]\end{aligned}\tag{24}$$

[0083] 其中，$(c_{1,l}^{k,\mathrm{RF}},c_{1,l}^{k,\mathrm{RF}})$表示第kth个混合森林组对新数据中第1个样本的增强映射，$(c_{n_{\mathrm{Raw}},l}^{k,\mathrm{RF}},c_{n_{\mathrm{Raw}},l}^{k,\mathrm{RF}})$表示第kth个混合森林组对新数据中第$n_{\mathrm{Raw}}$th个样本的增强映射，$(c_{N_{\mathrm{Raw}},l}^{k,\mathrm{RF}},c_{N_{\mathrm{Raw}},l}^{k,\mathrm{RF}})$表示第kth个混合森林组对新数据中第$N_{\mathrm{Raw}}$th个样本的增强映射；

[0084] 最后，特征增强层的输出$H^K$表示如下：

[0085] $$\boldsymbol{H}^{K}=[\boldsymbol{H}_{1},\boldsymbol{H}_{2},...,\boldsymbol{H}_{K}]\in R^{N_{\mathrm{Raw}}\times 2K} \tag{25}$$

[0086] 其中，$H_1$为第1个增强特征，$H_2$为第2个增强特征，$H_K$为第K个增强特征；

[0087] 当不考虑增量学习策略时，BHFR模型的表示如下：

[0088] $$\begin{aligned}\boldsymbol{Y}&=\boldsymbol{G}^{K}\boldsymbol{W}^{K}\\&=[\boldsymbol{Z}_{1},\boldsymbol{Z}_{2},...,\boldsymbol{Z}_{N}\,|\,\boldsymbol{H}_{1},\boldsymbol{H}_{2},...\boldsymbol{H}_{K}]\boldsymbol{W}^{K}\end{aligned} \tag{26}$$

[0089] 其中，$G^K$表示特征映射层与特征增强层输出的组合，即$G^K=[Z^N|H^K]$，其包含$N_{Raw}$个样本和(2N+2K)维特征；$W^K$表示特征映射层和特征增强层与输出层间的权重，其计算如下：

[0090] $$W^{K}=(\lambda I+[G^{K}]^{T}G^{K})^{-1}[G^{K}]^{T}Y \tag{27}$$

[0091] 其中，I表示单位矩阵，λ表示正则项系数；相应地，$G^K$的伪逆计算可表示为：

[0092] $$\begin{aligned}[\boldsymbol{G}^{K}]^{\dagger}&=(\lambda\boldsymbol{I}+[\boldsymbol{G}^{K}]^{\mathrm{T}}\boldsymbol{G})^{-1}[\boldsymbol{G}^{K}]^{\mathrm{T}}\\&=[\boldsymbol{Z}^{N}\,|\,\boldsymbol{H}^{K}]^{\dagger}\end{aligned} \tag{28}$$

[0093] 进一步的，所述步骤S4，构建增量学习层，通过增量式学习策略构建增量学习层，采用Moore-Penrose伪逆获得权重矩阵，进而实现BHFR软测量模型的高精度建模，具体包括：

[0094] 首先，对新数据集{X′,y}进行基于Bootstrap和RSM的采样，获取混合森林算法训练子集，过程如下：

[0095] $$\left\{\boldsymbol{X}'^{p,j}_{\mathrm{Bootstrap}},\boldsymbol{y}^{p,j}_{\mathrm{Bootstrap}}\right\}_{j=1}^{J}=\varphi_{p}^{\mathrm{ILL}}\left(\phi_{p}^{\mathrm{ILL}}\left\{\{\boldsymbol{X}',\boldsymbol{y}\},P_{\mathrm{Bootstrap}}\right\}\right) \tag{29}$$

[0096] 其中，$\boldsymbol{X}'^{p,j}_{\mathrm{Bootstrap}}$和$\boldsymbol{y}^{p,j}_{\mathrm{Bootstrap}}$为混合森林算法第个J训练子集的输入和输出，X′和y为新训练集的输入和输出，$\phi_i^{\mathrm{ILL}}(\cdot)$和$\varphi_p^{\mathrm{ILL}}(\cdot)$表示增量学习层中第pth个混合森林组的Bootstrap采样和RSM采样；

[0097] 接着，构建第pth个混合森林组中的决策树$f_{p,\mathrm{RF}}^{\mathrm{ILL}}(\cdot)$和$f_{p,\mathrm{CRF}}^{\mathrm{ILL}}(\cdot)$，其过程与特征映射层和特征增量层相同，此处不再赘述；

[0098] 进一步，当增加1个混合森林组后，特征映射层、特征增量层和增量学习层的输出$G^{K+1}$表示如下：

[0099] $$\begin{aligned}\boldsymbol{G}^{K+1}&=\left[\boldsymbol{G}^{K}\,|\,f_{1}^{\mathrm{FEL}}(\boldsymbol{X}')\right]\\&=\left[\boldsymbol{G}^{K}\,|\left\{f_{1,\mathrm{RF}}^{\mathrm{FEL}}(\boldsymbol{X}'),f_{1,\mathrm{CRF}}^{\mathrm{FEL}}(\boldsymbol{X}')\right\}\right]\\&=\left[\boldsymbol{G}^{K}\,|[(c_{1,l}^{1,\mathrm{RF}},c_{1,l}^{1,\mathrm{CRF}}),\ldots,(c_{N_{\mathrm{Raw}},l}^{1,\mathrm{RF}},c_{N_{\mathrm{Raw}},l}^{1,\mathrm{CRF}})]\right]\end{aligned} \tag{30}$$

[0100] 其中，$G^k=[Z^n|H^k]$包含$N_{Raw}$个样本和(2N+2K)维特征，$G^{K+1}$包含$N_{Raw}$个样本和(2N+2K+2J)维特征；

[0101] 然后，进行$G^{K+1}$的Moore-Penrose逆矩阵的递推更新，如下：

[0102] $$\boldsymbol{B}^{\mathrm{T}}=\begin{cases}[\boldsymbol{C}]^{\dagger}, & \text{if } \boldsymbol{C}\neq 0\\ [1+\boldsymbol{D}^{\mathrm{T}}\boldsymbol{D}]^{-1}\boldsymbol{D}^{\mathrm{T}}[\boldsymbol{G}^{k}]^{\dagger}, & \text{if } \boldsymbol{C}=0\end{cases} \tag{31}$$

[0103] 其中，矩阵C和矩阵D的计算如下：

[0104] $$\boldsymbol{C}=\boldsymbol{H}_{K+1}-\boldsymbol{G}^{K}\boldsymbol{D} \tag{32}$$

$$\boldsymbol{D}=[\boldsymbol{G}^{K}]^{\dagger}f_{1}^{\mathrm{ILL}}(\boldsymbol{X}^{\mathrm{New}}) \tag{33}$$

[0105] 进而，$G^{K+1}$的Moore-Penrose逆矩阵的递推公式如下：

[0106] $$[\boldsymbol{G}^{K+1}]^{\dagger}=\begin{bmatrix}[\boldsymbol{G}^{K}]^{\dagger}-\boldsymbol{D}\boldsymbol{B}^{\mathrm{T}}\\ \boldsymbol{B}^{\mathrm{T}}\end{bmatrix} \tag{34}$$

[0107] 进一步，计算特征映射层、特征增量层和增量学习层与输出层间权重的更新矩阵$W^{K+1}$，如下：

[0108] $$\boldsymbol{W}^{K+1}=\begin{bmatrix}\boldsymbol{W}^{K}-\boldsymbol{D}\boldsymbol{B}^{\mathrm{T}}\boldsymbol{Y}\\ \boldsymbol{B}^{\mathrm{T}}\boldsymbol{Y}\end{bmatrix} \tag{35}$$

[0109] 其中，$W^{K}=(\lambda I+[G^{K}]^{T}G^{K})^{-1}[G^{K}]^{T}Y$；

[0110] 由于采用上述伪逆更新策略只需要计算增量学习层混合森林组的伪逆矩阵，因此能够实现快速的增量式学习；

[0111] 进一步，根据训练误差的收敛程度实现自适应增量学习；

[0112] 定义误差的收敛阈值为$\theta_{\mathrm{Con}}$用以确定增量学习中混合森林组的数量p；相应地，BHFR模型的增量学习训练误差表示如下：

[0113] $$\ell=\lim_{p\to\infty}\left|\frac{1}{N}\left(\sqrt{\left(\boldsymbol{G}^{K+p}\boldsymbol{W}^{K+p}-\boldsymbol{y}\right)^{2}}-\sqrt{\left(\boldsymbol{G}^{K+p+1}\boldsymbol{W}^{K+p+1}-\boldsymbol{y}\right)^{2}}\right)\right|\leq\theta_{\mathrm{Con}} \tag{36}$$
$$\text{S.T.}\quad\theta_{\mathrm{Con}}\geq 0$$

[0114] 其中，1表示增量学习第p+1个与第p个混合森林组的训练误差值，$\sqrt{\left(\boldsymbol{G}^{K+p}\boldsymbol{W}^{K+p}-\boldsymbol{y}\right)^{2}}$和$\sqrt{\left(\boldsymbol{G}^{K+p+1}\boldsymbol{W}^{K+p+1}-\boldsymbol{y}\right)^{2}}$表示包含p个和p+1个混合森林组的BHFR模型训练误差；

[0115] 最终，所提BHFR软测量模型的预测输出$\hat{\boldsymbol{Y}}$为，

[0116] $$\hat{\boldsymbol{Y}}=\boldsymbol{G}^{K+P}\boldsymbol{W}^{K+P} \tag{37}$$

[0117] 根据本发明提供的具体实施例，本发明公开了以下技术效果：本发明提供的基于宽度混合森林回归的MSWI过程二噁英排放软测量方法，建立了基于BHFR的软测量模型，其结合了宽度学习建模、集成学习和潜在特征提取等算法，1)基于宽度学习系统框架，采用非微分学习器构建了包含特征映射层、潜在特征提取层、特征增强层和增量学习层的软测量模型；2)利用信息全联接、潜在特征提取和互信息度量对BHFR模型内部信息进行处理，有效保证了BHFR模型内部特征信息的传递最大化和冗余度最小化；3)采用混合森林组为映射单元实现建模过程的增量学习，通过伪逆策略快速计算输出层权重矩阵，再利用训练误差的收敛程度自适应调整增量学习，实现了高精度的软测量建模。在高维基准数据集和工业过程DXN数据集上验证了所提方法的有效性和合理性。

## 附图说明

[0118] 为了更清楚地说明本发明实施例或现有技术中的技术方案，下面将对实施例中所需要使用的附图作简单地介绍，显而易见地，下面描述中的附图仅仅是本发明的一些实施例，对于本领域普通技术人员来讲，在不付出创造性劳动性的前提下，还可以根据这些附图获得其他的附图。

[0119] 图1是本发明实施例基于宽度混合森林回归的MSWI过程二噁英排放软测量方法流程图；

[0120] 图2是本发明实施例城市固废焚烧过程工艺流程图；

[0121] 图3是本发明实施例训练误差收敛曲线；

[0122] 图4a是本发明实施例DXN数据集中训练集的拟合曲线；

[0123] 图4b是本发明实施例DXN数据集中验证集的拟合曲线；

[0124] 图4c是本发明实施例DXN数据集中测试集的拟合曲线。

## 具体实施方式

[0125] 下面将结合本发明实施例中的附图，对本发明实施例中的技术方案进行清楚、完整地描述，显然，所描述的实施例仅仅是本发明一部分实施例，而不是全部的实施例。基于本发明中的实施例，本领域普通技术人员在没有做出创造性劳动前提下所获得的所有其他实施例，都属于本发明保护的范围。

[0126] 本发明的目的是提供一种基于宽度混合森林回归的MSWI过程二噁英排放软测量方法，以MSWI过程DXN排放浓度检测为目标，提出了基于宽度混合森林回归(Broad HybridForest Regression,BHFR)的软测量建模算法。

[0127] 为使本发明的上述目的、特征和优点能够更加明显易懂，下面结合附图和具体实施方式对本发明作进一步详细的说明。

[0128] 如图1所示，本发明提供的基于宽度混合森林回归的MSWI过程二噁英排放软测量方法，包括如下步骤：

[0129] 基于BLS框架，以非微分基学习器替换神经元构建面向小样本高维数据的BHFR软测量模型，所述BHFR软测量模型包括特征映射层、潜在特征提取层、特征增强层和增量学习层的构建，具体包括以下步骤：

[0130] S1，构建特征映射层，构建由随机森林RF和完全随机森林CRF组成的混合森林组对高维特征进行映射；

[0131] S2，构建潜在特征提取层，依据贡献率对全联接混合矩阵的特征空间进行潜在特征提取，基于信息度量准则保证潜在有价值信息的最大化传递和最小化冗余，降低模型复杂度和计算消耗；

[0132] S3，构建特征增强层，基于所提取的潜在特征训练特征增强层以进一步增强特征表征能力；

[0133] S4，构建增量学习层，通过增量式学习策略构建增量学习层，采用Moore-Penrose伪逆获得权重矩阵，进而实现BHFR软测量模型的高精度建模；

[0134] S5，采用高维基准数据集和工业过程DXN数据集验证所述软测量模型；

[0135] S6，采用步骤S1-S5建立的软测量模型，对MSWI过程二噁英排放进行软测量。

[0136] MSWI过程包含固废储运、固废焚烧、余热锅炉、蒸汽发电、烟气净化和烟气排放等工艺阶段，以日处理量800吨的炉排式MSWI过程为例，其工艺流程如图2所示。

[0137] 结合DXN分解、生成、吸附和排放的全流程对各阶段的主要功能描述如下：

[0138] 1)固废储运阶段：环卫车辆从城市各收集站点将MSW运输至MSWI电厂，经称重记录后从卸料平台倾倒至固废储存池中未发酵区，然后由固废抓斗对其进行混合搅拌，再抓取至发酵区，经3～7天发酵和脱水以保证MSW焚烧的低位热值。研究表明，原生MSW中含有微量DXN(约0.8ng TEQ/Kg)，并含有DXN生成反应所需的多种含氯化合物。

[0139] 2)固废焚烧阶段:固废抓斗将发酵后的MSW投放至进料斗,经进料器将MSW推送到焚烧炉内,依次经过干燥、燃烧1、燃烧2和燃烬炉排后,MSW中的可燃成分随之完全燃烧;所需助燃空气由一次风机和二次风机从炉排下方和炉膛中部注入,最终燃烧产生的灰渣从燃烬炉排末端落至捞渣机,经水冷后送入炉渣池。为保证原生MSW中含有的以及焚烧时产生的DXN在炉内高温燃烧条件下能够被完全分解,炉膛燃烧过程需严格控制烟气温度在850℃以上、高温烟气在炉内停留时间超过2秒、确保足够大的烟气湍流度等工艺要求。

[0140] 3)余热锅炉阶段:炉膛产生的高温烟气(高于850℃)经引风机抽吸进入余热锅炉系统,先后经过过热器、蒸发器和省煤器设备,高温烟气与锅炉汽包液态水进行热交换后产生高温蒸汽,进而实现对高温烟气的降温处理,使余热锅炉出口的烟气温度低于200℃(即烟气G1)。从DXN生成机理的角度,高温烟气经余热锅炉降温时,导致DXN生成的化学反应包括高温气相合成反应(800℃～500℃)、前驱物合成(450℃～200℃)和从头合成(350℃～250℃)等,但目前还暂无统一的定论。

[0141] 4)蒸汽发电阶段:利用余热锅炉产生的高温蒸汽推动汽轮发电机,将机械能转变成电能,实现厂级用电的自给自足和剩余电量的上网供电,实现资源化和获取经济效益。

[0142] 5)烟气净化阶段:MSWI过程的烟气净化主要包含脱硝($NO_x$)、脱硫(HCL、HF、$SO_2$等)、脱重金属(Pb、Hg、Cd等)、吸附二噁英(DXN)和除尘(颗粒物)等一系列过程,进而实现焚烧烟气污染物排放达标的目的。采用活性炭喷射系统吸附焚烧烟气中DXN,是目前应用最广泛的技术手段,吸附后的DXN富集于飞灰中。

[0143] 6)烟气排放阶段:经降温和净化处理后的含有微量DXN的焚烧烟气(即烟气G2)由引风机抽吸经烟囱排放至大气中。MSWI过程的不间断、长时间的运行特性导致烟囱内壁颗粒物中附着大量DXN(即记忆效应),在何种工况下存在释放的可能性还是目前的研究难题。

[0144] 目前,面向MSWI过程的DXN软测量检测研究主要集中针对排放阶段(即烟气G3)的DXN浓度检测,本申请研究重点是构建G3烟气处的软测量模型。

[0145] 本申请所提BHFR建模策略包含特征映射层、潜在特征提取层、特征增强层和增量学习层四个主要部分。

[0146] 如图1中,$\{\boldsymbol{X},\boldsymbol{y}\}\in R^{N_{\mathrm{Raw}}\times(M+1)}$表示原始数据,其中$\boldsymbol{X}\in R^{N_{\mathrm{Raw}}\times M}$是原始输入数据,$N_{Raw}$是原始数据的数量,M是原始输入数据的维数,其来源于上述MSWI过程的六个不同阶段,以秒为单位在DCS系统采集与存储,$\boldsymbol{y}\in R^{N_{\mathrm{Raw}}\times 1}$是DXN排放浓度的输出真值,其来源于采用离线检测法得到排放物二噁英DXN检测样本;$\{DT_1,\cdots,DT_J\}$表示混合森林算法中的J个决策树模型,$DT_1$为第1个决策树模型,$DT_J$为第J个决策树模型;Bootstrap和RSM表示对输入数据进行样本和特征采样;$\{RF_n,CRF_n\}$表示第n个混合森林组模型,$RF_n$和$CRF_n$表示第n个RF和CRF模型;$\{\mathrm{Group}_n\}_{n=1}^{N}$表示特征映射层中包含N个混合森林组模型;$Z^N$表示特征映射层的输出;$H^K$表示特征增强层的输出;$[X|Z^N]$表示原始数据与$Z^N$的全联接混合矩阵;$\boldsymbol{X}'\in R^{N_{\mathrm{Raw}}\times M_{\mathrm{PCA}}}$表示经潜在特征提取后的新训练数据;$\{\mathrm{Group}_k\}_{k=1}^{K}$表示特征增强层包含的K个混合森林组模型;$\{\mathrm{Group}_p\}_{p=1}^{P}$表示增量学习层中包含的P个混合森林组模型;$W^{K+P}$表示最终的权重矩阵。

[0147] 各部分的主要功能如下:

[0148] 1)特征映射层:将来源于MSWI过程六个不同阶段的原始输入数据$\boldsymbol{X}\in R^{N_{\mathrm{Raw}}\times M}$通过特

征映射层的N个混合森林组$\{RF_n, CRF_n\}_{n=1}^{N}$进行特征映射，得到映射输出矩阵$Z^N$；

[0149] 2)潜在特征提取层：利用主成分分析对由原始输入数据$\boldsymbol{X} \in R^{N_{Raw} \times M}$与特征映射层输出$Z^N$组成的全联接混合矩阵[X|$Z^N$]进行潜在特征提取，去除特征空间的冗余信息，进一步通过所提取的潜在特征与DXN排放浓度的输出真值y的互信息确定潜在特征维数并得到新训练集$\boldsymbol{X}' \in R^{N_{Raw} \times M_{PM}}$；

[0150] 3)特征增强层：以新训练集$\boldsymbol{X}' \in R^{N_{Raw} \times M_{PM}}$作为输入，通过特征增强层的K个混合森林组$\{RF_k, CRF_k\}_{k=1}^{K}$组进行特征映射，得到增强层输出矩阵$H^K$；

[0151] 4)增量学习层：以新训练集$\boldsymbol{X}' \in R^{N_{Raw} \times M_{PM}}$作为输入，以混合森林组为最小单位逐步增加并更新权重$W^{K+P}$，直到训练误差收敛。

[0152] 从本质上讲，BHFR是以RF和CRF为基元构成的混合森林组作为基础映射单元取代原始BLS中的神经元；所述步骤S1，构建特征映射层，构建由随机森林RF和完全随机森林CRF组成的混合森林组对高维特征进行映射，具体包括：

[0153] 设原始数据为{X,y}，其中$\boldsymbol{X} \in R^{N_{Raw} \times M}$是原始输入数据，$N_{Raw}$是原始数据的数量，M是原始输入数据的维数，其来源于MSWI过程的六个不同阶段，以秒为单位在DCS系统采集与存储，$\boldsymbol{y} \in R^{N_{Raw} \times 1}$是DXN排放浓度的输出真值，其来源于采用离线检测法得到排放物DXN检测样本；以特征映射层的第nth个混合森林组为例描述特征映射层的建模过程：

[0154] 对{X,y}进行Bootstrap和随机子空间RSM采样，获得混合森林组模型的J个训练子集，如下：

[0155] $$\left\{\boldsymbol{X}_{\text{Bootstrap}}^{n,j}, \boldsymbol{y}_{\text{Bootstrap}}^{n,j}\right\}_{j=1}^{J} = \varphi_n^{\text{FML}}\left(\phi_n^{\text{FML}}((\boldsymbol{X},\boldsymbol{y}), P_{\text{Bootstrap}})\right) \quad (1)$$

[0156] 其中，$\boldsymbol{X}_{\text{Bootstrap}}^{n,j}$和$\boldsymbol{y}_{\text{Bootstrap}}^{n,j}$为第J个训练子集的输入和输出，$\phi_n^{\text{FML}}(\cdot)$和$\varphi_n^{\text{FML}}(\cdot)$表示特征映射层中对第nth个混合森林组的Bootstrap和RSM采样，$P_{\text{Bootstrap}}$表示Bootstrap采样概率；

[0157] 基于$\left\{\boldsymbol{X}_{\text{Bootstrap}}^{n,j}, \boldsymbol{y}_{\text{Bootstrap}}^{n,j}\right\}_{j=1}^{J}$训练包含J个决策树的混合森林算法，其中特征映射层中的第nth个混合森林组的第jth个决策树表示如下：

[0158] $$f_{n,j}^{\text{DT}}(\cdot) = \sum_{l=1}^{L} c_l I(\boldsymbol{x}_{\text{Bootstrap}}^{n,j} \in R_l), \quad l = 1,2,...,L \quad (2)$$

[0159] 其中，L表示决策树叶节点数量，I(•)表示指示函数，$c_1$采用递归分裂方式计算；

[0160] RF中决策树的分裂损失函数$\Omega_i(\cdot)$表示为：

[0161] $$\begin{aligned}\Omega_i(s,v) &= \min\left([\boldsymbol{y}_L - E[\boldsymbol{y}_L]] + [\boldsymbol{y}_R - E[\boldsymbol{y}_R]]\right) \\ &= \min\left(\sum_{x_{\text{Bootstrap}}^{n,j} \in R_L} (y_L^i - c_L)^2 + \sum_{x_{\text{Bootstrap}}^{n,j} \in R_R} (y_R^i - c_R)^2\right)\end{aligned} \quad (3)$$

[0162] 其中，$\Omega_i(s,v)$表示第sth个特征的值v作为切分准则的损失函数值，$y_L$表示左叶节点的DXN排放浓度真值向量，E[$y_L$]表示$y_L$的数学期望，$y_R$表示右叶节点的DXN排放浓度真值向量，E[$y_R$]表示$y_R$的数学期望，$y_L^i$表示左叶节点第i个DXN排放浓度真值，$y_R^i$表示右叶节点第i个DXN排放浓度真值，$c_L$表示左叶节点DXN排放浓度预测输出，$c_R$表示右叶节点DXN排放浓度预测输出；

[0163] 通过最小化$\Omega_i(s,v)$，将训练集$(\boldsymbol{X}_{\text{Bootstrap}}^{n,j},\boldsymbol{y}_{\text{Bootstrap}}^{n,j})$切分为两个树节点，如下：

[0164] $$\min\left\{\Omega_i(s,v)\right\}_{i=1}^{N_{\text{Raw}}\times M}\xrightarrow{\text{树节点分裂}}\begin{cases}R_{\text{L}}^{N_{\text{L}}\times M}\\R_{\text{R}}^{N_{\text{R}}\times M}\end{cases}\tag{4}$$

[0165] 其中，$R_{\text{L}}^{N_{\text{L}}\times M}$和$R_{\text{R}}^{N_{\text{R}}\times M}$表示切分后左右两个树节点所包含的样本集，$N_L$和$N_R$分别表示$R_{\text{L}}^{N_{\text{L}}\times M}$和$R_{\text{R}}^{N_{\text{R}}\times M}$中的样本数量；

[0166] 当前左右树节点的DXN排放浓度预测输出值输出值$c_{\text{L}}^{\text{RF}}$和$c_{\text{R}}^{\text{RF}}$为样本真值的期望，如下：

[0167] $$\begin{cases}c_{\text{L}}^{\text{RF}}=E[\boldsymbol{y}_{\text{L}}],\ \ \boldsymbol{y}_{\text{L}}\in R_{\text{L}}^{N_{\text{L}}\times M}\\c_{\text{R}}^{\text{RF}}=E[\boldsymbol{y}_{\text{R}}],\ \ \boldsymbol{y}_{\text{R}}\in R_{\text{R}}^{N_{\text{R}}\times M}\end{cases}\tag{5}$$

[0168] 其中，$y_L$和$y_R$表示$R_{\text{L}}^{N_{\text{L}}\times M}$和$R_{\text{R}}^{N_{\text{R}}\times M}$中的DXN排放浓度真值向量，E[$y_L$]和E[$y_R$]表示$y_L$和$y_R$的数学期望；

[0169] 与RF不同，CRF中决策树分裂采用完全随机选择方式，表示为，

[0170] $$rand\left\{(s,v)_i\right\}_{i=1}^{N_{\text{Raw}}\times M}\xrightarrow{\text{树节点分裂}}\begin{cases}R_{\text{L}}^{N_{\text{L}}\times M}\\R_{\text{R}}^{N_{\text{R}}\times M}\end{cases}\tag{6}$$

[0171] 其中，$rand\left\{(s,v)_i\right\}_{i=1}^{N_{\text{Raw}}\times M}$表示完全随机选取第sth个特征的值v作为切分点；

[0172] 被随机分裂的左右树节点的DXN排放浓度预测输出值$c_{\text{L}}^{\text{CRF}}$和$c_{\text{R}}^{\text{CRF}}$为样本真值的期望，如下：

[0173] $$\begin{aligned}c_{\text{L}}^{\text{CRF}}&=E[\boldsymbol{y}_{\text{L}}],\ \ \boldsymbol{y}_{\text{L}}\in R_{\text{L}}^{N_{\text{L}}\times M}\\c_{\text{R}}^{\text{CRF}}&=E[\boldsymbol{y}_{\text{R}}],\ \ \boldsymbol{y}_{\text{R}}\in R_{\text{R}}^{N_{\text{R}}\times M}\end{aligned}\tag{7}$$

[0174] 通过上述过程，第nth个混合森林组$f_n^{\text{FML}}(\cdot)$可表示为，

[0175] $$f_n^{\text{FML}}(\cdot)=\left\{f_{n,\text{RF}}^{\text{FML}}(\cdot),f_{n,\text{CRF}}^{\text{FML}}(\cdot)\right\}\tag{8}$$

[0176] 其中，$f_{n,\text{RF}}^{\text{FML}}(\cdot)$表示第nth个随机森林，$f_{n,\text{CRF}}^{\text{FML}}(\cdot)$表示第nth个完全随机森林；进而，第nth个映射特征$Z_n$可表示为

[0177] $$\begin{aligned}\boldsymbol{Z}_n&=f_n^{\text{FML}}(\boldsymbol{X})=\left\{f_{n,\text{RF}}^{\text{FML}}(\boldsymbol{X}),f_{n,\text{CRF}}^{\text{FML}}(\boldsymbol{X})\right\}\\&=[(c_{1,l}^{n,\text{RF}},c_{1,l}^{n,\text{RF}}),\ldots,(c_{n_{\text{Raw}},l}^{n,\text{RF}},c_{n_{\text{Raw}},l}^{n,\text{RF}}),\ldots,(c_{N_{\text{Raw}},l}^{n,\text{RF}},c_{N_{\text{Raw}},l}^{n,\text{RF}})]\end{aligned}\tag{9}$$

[0178] 其中，$(c_{1,l}^{n,\text{RF}},c_{1,l}^{n,\text{RF}})$表示第nth组混合森林对来源于MSWI过程六个不同阶段的原始输入数据第1个样本的映射特征，$(c_{n_{\text{Raw}},l}^{n,\text{RF}},c_{n_{\text{Raw}},l}^{n,\text{RF}})$表示第nth组混合森林对来源于MSWI过程六个不同阶段的原始输入数据第$n_{\text{Raw}}$th个样本的映射特征，$(c_{N_{\text{Raw}},l}^{n,\text{RF}},c_{N_{\text{Raw}},l}^{n,\text{RF}})$表示第nth组混合森林对来源于MSWI过程六个不同阶段的原始输入数据第$N_{\text{Raw}}$th个样本的映射特征；

[0179] 最终，特征映射层的输出表示为：

[0180] $$\boldsymbol{Z}^N=\left(\boldsymbol{Z}_1,\boldsymbol{Z}_2,\ldots,\boldsymbol{Z}_N\right)\in R^{N_{\text{Raw}}\times 2N}\tag{10}$$

[0181] 其中，$Z_1$为第1个映射特征，$Z_2$为第2个映射特征，$Z_N$为第N个映射特征，映射特征矩

阵$Z^N$包含$N_{Raw}$个样本和2N维特征。

[0182] 为了避免信息传递过程中的信息丢失导致的过拟合现象，本申请所提BHFR采用全联接策略实现特征映射层与特征增强层、增量学习层之间的信息传递。同时，为了保证模型训练过程中信息冗余最小化，此处采用主成分分析(Principal ComponentAnalysis，PCA)提取全联接混合矩阵特征空间的潜在特征，再利用互信息进一步筛选与真值信息最大化相关的潜在特征，进而实现对高维数据的降维处理。

[0183] 所述步骤S2，构建潜在特征提取层，依据贡献率对全联接混合矩阵的特征空间进行潜在特征提取，基于信息度量准则保证潜在有价值信息的最大化传递和最小化冗余，降低模型复杂度和计算消耗，具体包括：

[0184] 首先，来源于MSWI过程六个不同阶段的原始输入数据X与特征映射矩阵$Z^N$组合得到全联接混合矩阵A，表示为：

[0185] $$\boldsymbol{A}=[\boldsymbol{X}\mid\boldsymbol{Z}^N]\in R^{N_{\mathrm{Raw}}\times(M+2N)} \tag{11}$$

[0186] 其中，A含$N_{Raw}$个样本和(M+2N)维特征；

[0187] 接着，考虑到A的维数远高于原始数据，此处利用PCA最小化A中的冗余信息，计算A的相关矩阵R，如下：

[0188] $$\boldsymbol{R}=\frac{1}{N_{\mathrm{Raw}}-1}\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}\in R^{(M+2N)\times(M+2N)} \tag{12}$$

[0189] 进一步，对R进行奇异值分解，得到(M+2N)个特征值和相应特征向量，如下：

[0190] $$R=U_{(M+2N)}\Sigma_{(M+2N)}V_{(M+2N)} \tag{13}$$

[0191] 其中，$U_{(M+2N)}$表示(M+2N)阶正交矩阵，$\Sigma_{(M+2N)}$表示(M+2N)阶对角矩阵，$V_{(M+2N)}$表示(M+2N)阶正交矩阵；

[0192] $$\boldsymbol{\Sigma}_{(M+2N)}=\begin{bmatrix}\sigma_1 & & \\ & \ddots & \\ & & \sigma_{(M+2N)}\end{bmatrix} \tag{14}$$

[0193] 其中，$\sigma_1>\sigma_2>\cdots>\sigma_{(M+2N)}$表示由大到小排列的特征值；

[0194] 然后，根据设定潜在特征贡献阈值η，确定最终的主成分数量，

[0195] $$\eta=\sum_{q=1}^{Q_{\mathrm{PCA}}}\sigma_q\Big/\sum_{q=1}^{(M+2N)}\sigma_q \tag{15}$$

[0196] 其中，潜在特征数量$Q_{PCA}$<<(M+2N)；

[0197] 基于上述确定的$Q_{PCA}$个潜在特征，获得特征值集合$\left\{\sigma_q\right\}_{q=1}^{Q_{\mathrm{PCA}}}$对应的特征向量矩阵$V_{QPCA}$，即A的投影矩阵；然后，对A进行特征投影以实现冗余信息的最小化处理，将获得潜在特征记为$X^{PCA}$，即

[0198] $$\boldsymbol{X}^{\mathrm{PCA}}=\boldsymbol{A}\boldsymbol{V}_{I_{\mathrm{PCA}}}\in R^{N_{\mathrm{Raw}}\times M_{\mathrm{PCA}}} \tag{16}$$

[0199] 其中，$\boldsymbol{V}_{Q_{\mathrm{PCA}}}\in R^{(M+2N)\times Q_{\mathrm{PCA}}}$表示前$Q_{PCA}$个潜在特征的特征向量；

[0200] 进一步，计算所选潜在特征$X^{PCA}$与真值$\boldsymbol{y}\in\boldsymbol{R}^{N_{\mathrm{Raw}}\times 1}$间的互信息值$I^{MI}$，如下：

[0201] $$I^{\mathrm{MI}}(\boldsymbol{X}^{\mathrm{PCA}},\boldsymbol{y})=\sum_{q=1}^{Q_{\mathrm{PCA}}}p(\boldsymbol{x}_q^{\mathrm{PCA}},\boldsymbol{y})\log_2\frac{p(\boldsymbol{x}_q^{\mathrm{PCA}},\boldsymbol{y})}{p(\boldsymbol{x}_q^{\mathrm{PCA}}),p(\boldsymbol{y})} \tag{17}$$

[0202] 其中，$p(\boldsymbol{x}_q^{\mathrm{PCA}},\boldsymbol{y})$表示第qth个潜在特征$\boldsymbol{x}_q^{\mathrm{PCA}}$与DXN排放浓度真值y的联合概率分布，$p(\boldsymbol{x}_q^{\mathrm{PCA}})$表示第qth个潜在特征$\boldsymbol{x}_q^{\mathrm{PCA}}$的边缘概率分布，p(y)表示DXN排放浓度真值y的边缘概率分布；

[0203] 接着，通过信息最大化选择机制以保证所选择潜在特征与真值的相关性，表示为：

[0204] $$\left\{I_q^{\mathrm{MI}}\right\}_{q=1}^{Q_{\mathrm{PCA}}}\xrightarrow{I_q^{\mathrm{MI}}\geq\zeta}\left\{I_q^{\mathrm{MI}}\right\}_{q=1}^{Q_{\mathrm{PCA}}^{\mathrm{MI}}} \tag{18}$$

[0205] 其中，$\left\{I_q^{\mathrm{MI}}\right\}_{q=1}^{Q}$表示$Q_{\mathrm{PCA}}$个潜在特征$\boldsymbol{x}_q^{\mathrm{PCA}}$与真值y的互信息值，ζ表示最大化信息的阈值，$\left\{I_q^{\mathrm{MI}}\right\}_{q=1}^{Q_{\mathrm{PCA}}^{\mathrm{MI}}}$表示与DXN排放浓度真值y信息相关度最大的$Q_{\mathrm{PCA}}^{\mathrm{MI}}$个潜在特征；

[0206] 最终，获得包括$Q_{\mathrm{PCA}}^{\mathrm{MI}}$个潜在特征的新数据集$\{\boldsymbol{X}',\boldsymbol{y}\}\in R^{N_{\mathrm{Raw}}\times\left(Q_{\mathrm{PCA}}^{\mathrm{MI}}+1\right)}$，并设定提取后维数$M_{\mathrm{PCA}}^{\mathrm{MI}}=Q_{\mathrm{PCA}}^{\mathrm{MI}}$。

[0207] 所述步骤S3中，构建特征增强层，基于所提取的潜在特征训练特征增强层以进一步增强特征表征能力，具体包括：

[0208] 首先对新数据集{X′,y}进行基于Bootstrap和RSM的采样，获取混合森林算法的第个J训练子集，如下：

[0209] $$\left\{\boldsymbol{X}'^{k,j}_{\mathrm{Bootstrap}},\boldsymbol{y}^{k,j}_{\mathrm{Bootstrap}}\right\}_{j=1}^{J}=\varphi_k^{\mathrm{FEL}}\left(\phi_k^{\mathrm{FEL}}\left(\{\boldsymbol{X}',\boldsymbol{y}\},P_{\mathrm{Bootstrap}}\right)\right) \tag{19}$$

[0210] 其中，$\boldsymbol{X}'^{k,j}_{\mathrm{Bootstrap}}$和$\boldsymbol{y}^{k,j}_{\mathrm{Bootstrap}}$为第个J训练子集的输入和输出，X′和y为新训练集的输入和输出，$\phi_k^{\mathrm{FEL}}(\cdot)$表示对第kth个混合森林组的Bootstrap采样，$\varphi_k^{\mathrm{FEL}}(\cdot)$表示对第kth个混合森林组的RSM采样；

[0211] 接着，以第kth个混合森林组中第j个RF的构建为例，如下：

[0212] $$\left\{\boldsymbol{X}'^{k,j}_{\mathrm{Bootstrap}},\boldsymbol{y}^{k,j}_{\mathrm{Bootstrap}}\right\}\xrightarrow{\Omega_j(s,v)}f_{k,j}^{\mathrm{DT\text{-}RF}}(\cdot)=\sum_{l=1}^{L}c_lI(\boldsymbol{x}^{k,j}_{\mathrm{Bootstrap}}\in R_l),\quad l=1,2,...,L \tag{20}$$

[0213] 其中，$f_{k,j}^{\mathrm{DT\text{-}RF}}(\cdot)$表示特征增强层中第kth个混合森林组中RF的第jth个决策树；L表示决策树叶节点的数量；$c_1$采用递归分裂方式计算，具体过程公式(3)-(5)；

[0214] 进而，可得到特征增强层中第kth个混合森林组中的RF模型，其表示为，

[0215] $$f_{k,\mathrm{RF}}^{\mathrm{FEL}}(\cdot)=\left\{f_{k,j}^{\mathrm{DT\text{-}RF}}(\cdot)\right\}_{j=1}^{J} \tag{21}$$

[0216] 然后，类似地以第kth个混合森林组中的第j个CRF的构建为例，如下：

[0217] $$\left\{\boldsymbol{X}'^{k,j}_{\mathrm{Bootstrap}},\boldsymbol{y}^{k,j}_{\mathrm{Bootstrap}}\right\}\xrightarrow{rand_j(s,v)}f_{k,j}^{\mathrm{DT\text{-}CRF}}(\cdot)=\sum_{l=1}^{L}c_lI(\boldsymbol{x}^{k,j}_{\mathrm{Bootstrap}}\in R_l),\quad l=1,2,...,L \tag{22}$$

[0218] 其中，$f_{k,j}^{\mathrm{DT\text{-}CRF}}(\cdot)$表示特征增强层中第kth个混合森林组中CRF的第jth个决策树；$c_1$采用递归分裂方式计算，具体过程见公式(6)-(7)；

[0219] 进而，可得到特征增强层中第kth个混合森林组的CRF模型，其表示为，

[0220] $$f_{k,\mathrm{CRF}}^{\mathrm{FEL}}(\cdot)=\left\{f_{k,j}^{\text{DT-CRF}}(\cdot)\right\}_{j=1}^{J} \tag{23}$$

[0221] 通过上述过程，得到第kth个混合森林组$f_k^{\mathrm{FEL}}(\cdot)$；进而，第kth个增强特征可表示如下：

[0222] $$\begin{aligned}\boldsymbol{H}_k&=f_k^{\mathrm{FEL}}(\boldsymbol{X}')=\left[f_{k,\mathrm{RF}}^{\mathrm{FEL}}(\boldsymbol{X}'),f_{k,\mathrm{CRF}}^{\mathrm{FEL}}(\boldsymbol{X}')\right]\\&=\left[(c_{1,l}^{k,\mathrm{RF}},c_{1,l}^{k,\mathrm{CRF}}),\ldots,(c_{n_{\mathrm{Raw}},l}^{k,\mathrm{RF}},c_{n_{\mathrm{Raw}},l}^{k,\mathrm{RF}}),\ldots,(c_{N_{\mathrm{Raw}},l}^{k,\mathrm{RF}},c_{N_{\mathrm{Raw}},l}^{k,\mathrm{CRF}})\right]\end{aligned} \tag{24}$$

[0223] 其中，$(c_{1,l}^{k,\mathrm{RF}},c_{1,l}^{k,\mathrm{RF}})$表示第kth个混合森林组对新数据中第1个样本的增强映射，$(c_{n_{\mathrm{Raw}},l}^{k,\mathrm{RF}},c_{n_{\mathrm{Raw}},l}^{k,\mathrm{RF}})$表示第kth个混合森林组对新数据中第$n_{\mathrm{Raw}}$th个样本的增强映射，$(c_{N_{\mathrm{Raw}},l}^{k,\mathrm{RF}},c_{N_{\mathrm{Raw}},l}^{k,\mathrm{RF}})$表示第kth个混合森林组对新数据中第$N_{\mathrm{Raw}}$th个样本的增强映射；

[0224] 最后，特征增强层的输出$H^K$表示如下：

[0225] $$\boldsymbol{H}^K=[\boldsymbol{H}_1,\boldsymbol{H}_2,\ldots,\boldsymbol{H}_K]\in R^{N_{\mathrm{Raw}}\times 2K} \tag{25}$$

[0226] 其中，$H_1$为第1个增强特征，$H_2$为第2个增强特征，$H_K$为第K个增强特征；

[0227] 当不考虑增量学习策略时，BHFR模型的表示如下：

[0228] $$\begin{aligned}\boldsymbol{Y}&=\boldsymbol{G}^K\boldsymbol{W}^K\\&=[\boldsymbol{Z}_1,\boldsymbol{Z}_2,\ldots,\boldsymbol{Z}_N\,|\,\boldsymbol{H}_1,\boldsymbol{H}_2,\ldots\boldsymbol{H}_K]\boldsymbol{W}^K\end{aligned} \tag{26}$$

[0229] 其中，$G^K$表示特征映射层与特征增强层输出的组合，即$G^K=[Z^N|H^K]$，其包含$N_{\mathrm{Raw}}$个样本和(2N+2K)维特征；$W^K$表示特征映射层和特征增强层与输出层间的权重，其计算如下：

[0230] $$W^K=(\lambda I+[G^K]^T G^K)^{-1}[G^K]^T Y \tag{27}$$

[0231] 其中，I表示单位矩阵，λ表示正则项系数；相应地，$G^K$的伪逆计算可表示为：

[0232] $$\begin{aligned}[\boldsymbol{G}^K]^{\dagger}&=(\lambda\boldsymbol{I}+[\boldsymbol{G}^K]^{\mathrm{T}}\boldsymbol{G})^{-1}[\boldsymbol{G}^K]^{\mathrm{T}}\\&=[\boldsymbol{Z}^N\,|\,\boldsymbol{H}^K]^{\dagger}\end{aligned} \tag{28}$$

[0233] 本申请所提的BHFR以混合森林组为基本单元依据训练误差的收敛程度实现增量学习。所述步骤S4，构建增量学习层，通过增量式学习策略构建增量学习层，采用Moore-Penrose伪逆获得权重矩阵，进而实现BHFR软测量模型的高精度建模，具体包括：

[0234] 首先，对新数据集{X′,y}进行基于Bootstrap和RSM的采样，获取混合森林算法训练子集，过程如下：

[0235] $$\left\{\boldsymbol{X}'^{p,j}_{\mathrm{Bootstrap}},\boldsymbol{y}^{p,j}_{\mathrm{Bootstrap}}\right\}_{j=1}^{J}=\varphi_p^{\mathrm{ILL}}\left(\phi_p^{\mathrm{ILL}}\left\{\{\boldsymbol{X}',\boldsymbol{y}\},P_{\mathrm{Bootstrap}}\right\}\right) \tag{29}$$

[0236] 其中，$\boldsymbol{X}'^{p,j}_{\mathrm{Bootstrap}}$和$\boldsymbol{y}^{p,j}_{\mathrm{Bootstrap}}$为混合森林算法第个J训练子集的输入和输出，X′和y为新训练集的输入和输出，$\phi_i^{\mathrm{ILL}}(\cdot)$和$\varphi_p^{\mathrm{ILL}}(\cdot)$表示增量学习层中第pth个混合森林组的Bootstrap采样和RSM采样；

[0237] 接着，构建第pth个混合森林组中的决策树$f_{p,\mathrm{RF}}^{\mathrm{ILL}}(\cdot)$和$f_{p,\mathrm{CRF}}^{\mathrm{ILL}}(\cdot)$，其过程与特征映射层和特征增量层相同，此处不再赘述；

[0238] 进一步，当增加1个混合森林组后，特征映射层、特征增量层和增量学习层的输出$G^{K+1}$表示如下：

$$\begin{aligned} \boldsymbol{G}^{K+1} &= \left[\boldsymbol{G}^{K} \mid f_1^{\mathrm{FEL}}(\boldsymbol{X}')\right] \\ &= \left[\boldsymbol{G}^{K} \mid \left\{f_{1,\mathrm{RF}}^{\mathrm{FEL}}(\boldsymbol{X}'), f_{1,\mathrm{CRF}}^{\mathrm{FEL}}(\boldsymbol{X}')\right\}\right] \\ &= \left[\boldsymbol{G}^{K} \mid [(c_{1,J}^{1,\mathrm{RF}}, c_{1,J}^{1,\mathrm{CRF}}), \ldots, (c_{N_{\mathrm{Raw}},J}^{1,\mathrm{RF}}, c_{N_{\mathrm{Raw}},J}^{1,\mathrm{CRF}})]\right] \end{aligned} \tag{30}$$

[0240] 其中，$G^k=[Z^n|H^k]$包含$N_{Raw}$个样本和(2N+2K)维特征，$G^{K+1}$包含$N_{Raw}$个样本和(2N+2K+2J)维特征；

[0241] 然后，进行$G^{K+1}$的Moore-Penrose逆矩阵的递推更新，如下：

$$\boldsymbol{B}^{\mathrm{T}} = \begin{cases} [\boldsymbol{C}]^{\dagger}, & \text{if } \boldsymbol{C} \neq 0 \\ [1+\boldsymbol{D}^{\mathrm{T}}\boldsymbol{D}]^{-1}\boldsymbol{D}^{\mathrm{T}}[\boldsymbol{G}^{k}]^{\dagger}, & \text{if } \boldsymbol{C} = 0 \end{cases} \tag{31}$$

[0243] 其中，矩阵C和矩阵D的计算如下：

[0244] $C=H_{K+1}-G^KD$ (32)

$$\boldsymbol{D} = [\boldsymbol{G}^{K}]^{\dagger} f_1^{\mathrm{ILL}}(\boldsymbol{X}^{\mathrm{New}}) \tag{33}$$

[0246] 进而，$G^{K+1}$的Moore-Penrose逆矩阵的递推公式如下：

$$[\boldsymbol{G}^{K+1}]^{\dagger} = \begin{bmatrix} [\boldsymbol{G}^{K}]^{\dagger} - \boldsymbol{D}\boldsymbol{B}^{\mathrm{T}} \\ \boldsymbol{B}^{\mathrm{T}} \end{bmatrix} \tag{34}$$

[0248] 进一步，计算特征映射层、特征增量层和增量学习层与输出层间权重的更新矩阵$W^{K+1}$，如下：

$$\boldsymbol{W}^{K+1} = \begin{bmatrix} \boldsymbol{W}^{K} - \boldsymbol{D}\boldsymbol{B}^{\mathrm{T}}\boldsymbol{Y} \\ \boldsymbol{B}^{\mathrm{T}}\boldsymbol{Y} \end{bmatrix} \tag{35}$$

[0250] 其中，$W^K=(\lambda I+[G^K]^TG^K)^{-1}[G^K]^TY$；

[0251] 由于采用上述伪逆更新策略只需要计算增量学习层混合森林组的伪逆矩阵，因此能够实现快速的增量式学习；

[0252] 进一步，根据训练误差的收敛程度实现自适应增量学习；

[0253] 定义误差的收敛阈值为$\theta_{Con}$用以确定增量学习中混合森林组的数量p；相应地，BHFR模型的增量学习训练误差表示如下：

$$\ell = \lim_{p\to\infty}\left|\frac{1}{N}\left(\sqrt{\left(\boldsymbol{G}^{K+p}\boldsymbol{W}^{K+p}-\boldsymbol{y}\right)^2} - \sqrt{\left(\boldsymbol{G}^{K+p+1}\boldsymbol{W}^{K+p+1}-\boldsymbol{y}\right)^2}\right)\right| \le \theta_{\mathrm{Con}} \quad \text{S.T. } \theta_{\mathrm{Con}} \ge 0 \tag{36}$$

[0255] 其中，$\ell$表示增量学习第p+1个与第p个混合森林组的训练误差值，$\sqrt{\left(\boldsymbol{G}^{K+p}\boldsymbol{W}^{K+p}-\boldsymbol{y}\right)^2}$和$\sqrt{\left(\boldsymbol{G}^{K+p+1}\boldsymbol{W}^{K+p+1}-\boldsymbol{y}\right)^2}$表示包含p个和p+1个混合森林组的BHFR模型训练误差；

[0256] 最终，所提BHFR软测量模型的预测输出$\hat{\boldsymbol{Y}}$为，

$$\hat{\boldsymbol{Y}} = \boldsymbol{G}^{K+P}\boldsymbol{W}^{K+P} \tag{37}$$

[0258] 本申请采用某MSWI电厂的实际DXN数据进行工业验证。DXN数据源自于北京某MSWI焚烧发电厂，共涵盖了2009-2020年的DXN排放浓度建模数据141组，DXN真值为2小时采样化验后的折算浓度，对缺失数据和异常变量进行剔除后的输入变量为116维，相应地取值为当前DXN真值采样时间段内的均值。

[0259] 本申请选取均方根误差(Root Mean Square Error,RMSE)、平均绝对误差MAE和决定系数(Coefficient ofDetermination,$R^2$)共三个评价指标比较不同方法的性能,计算如下:

[0260] $$RMSE=\sqrt{\sum_{i=1}^{N}(y_i-\hat{y}_i)^2/(N-1)} \tag{38}$$

[0261] $$MAE=\frac{1}{N}\sum_{i=1}^{N}|\hat{y}_i-y_i| \tag{39}$$

[0262] $$R^2=1-\sum_{i=1}^{N}(y_i-\hat{y}_i)^2\Big/\sum_{i=1}^{N}(y_i-\bar{y})^2 \tag{40}$$

[0263] 其中,N为数据的数量,$y_i$为第i个真值,$\hat{y}_i$为第i个预测值,$\bar{y}$为均值。

[0264] 在DXN数据集中,BHFR方法的参数设置为:决策树叶节点最小样本数$N_{smples}$为7,RSM特征选择数量$\sqrt{N_{features}}$,决策树的数量$N_{tree}$为10,特征映射层和特征增强层中混合森林组的数量$N_{Forest}$均为10,潜在特征贡献率阈值η为0.9,正则化参数λ为2^-10。

[0265] 类似基准数据集,首先基于全联接混合矩阵和特征空间A确定用于特征增强层和增量学习层潜在特征数量。在DXN数据集中A的特征维数为316维。当潜在特征贡献率阈值η为0.9时,DXN数据集中选择的潜在特征数量分别为35个。接着,计算35个潜在特征与DXN真值间的互信息值。将互信息阈值ζ设置为0.75,DXN数据集中被选的潜在特征数量为6个。

[0266] 进一步,预设增量学习层的混合森林组单元数量为1000,相应地BHFR模型的训练误差与混合森林组数量间的关系如图3所示。

[0267] 由图3所示的训练误差曲线可知,BHFR在DXN数据集上的训练过程可收敛至某一确定下限值。

[0268] 然后,采用RF、DFR、DFR-clfc和BLS-NN与所提BHFR进行对比,参数设置为:(1)RF,决策树叶节点最小样本数$N_{smples}$为3,RSM特征选择数量为$\sqrt{N_{features}}$,决策树的数量$N_{tree}$为500;(2)DFR,决策树叶节点最小样本数$N_{smples}$为3,RSM特征选择数量为$\sqrt{N_{features}}$,决策树的数量$N_{tree}$为500,每层中RF和CRF模型的数量$N_{RF}$和$N_{CRF}$均为2,总层数设置为50;(3)DFR-clfc,决策树叶节点最小样本数$N_{smples}$为3,RSM特征选择数量为$\sqrt{N_{features}}$,决策树的数量$N_{tree}$为500,每层中RF和CRF模型的数量$N_{RF}$和$N_{CRF}$均为2,总层数设置为50;(4)BLS-NN,特征节点数$N_m$为5,增强节点数$N_e$为41,神经元数量$N_n$为9和正则化参数λ为2^30。上述方法在相同条件下重复20次实验,其统计结果和预测曲线如表1和图4a-4c所示。

[0269] 表1 DXN数据集实验结果

[0270]

| 方法 | 数据集 | RMSE | | MAE | | R2 | |
|---|---|---|---|---|---|---|---|
| | | 平均值 | 方差 | 平均值 | 方差 | 平均值 | 方差 |
| RF | 训练集 | 1.1159E-02 | 5.7497E-08 | 9.0221E-03 | 4.0684E-08 | 8.5346E-01 | 3.9360E-05 |
| | 验证集 | **2.0051E-02** | 1.8026E-07 | 1.4677E-02 | 8.2255E-08 | **5.0196E-01** | 4.3515E-04 |
| | 测试集 | 1.6922E-02 | 1.6150E-07 | 1.3548E-02 | 8.9520E-08 | 5.9001E-01 | 3.7817E-04 |
| DFR | 训练集 | 1.1493E-02 | 8.7413E-09 | 9.4568E-03 | 4.6626E-09 | 8.4463E-01 | 6.3663E-06 |
| | 验证集 | 2.0735E-02 | **9.7835E-09** | 1.5780E-02 | **1.1121E-08** | 4.6759E-01 | **2.5813E-05** |
| | 测试集 | 1.7791E-02 | **1.7308E-08** | 1.4608E-02 | 1.5235E-08 | 5.4701E-01 | **4.5066E-05** |
| DFR-clfc | 训练集 | 8.0852E-03 | 2.9078E-06 | 6.6040E-03 | 2.0819E-06 | 9.1986E-01 | 1.1887E-03 |
| | 验证集 | 2.0187E-02 | 1.4562E-07 | 1.5626E-02 | 2.3355E-08 | 4.9520E-01 | 3.6404E-04 |
| | 测试集 | 1.7025E-02 | 1.5755E-07 | 1.4068E-02 | 6.0233E-08 | 5.8501E-01 | 3.7843E-04 |
| BLS-NN | 训练集 | **1.2924E-09** | **1.5756E-18** | **9.5358E-10** | **7.2150E-19** | **1.0000E+00** | **8.2358E-29** |
| | 验证集 | 6.8845E-02 | 7.0040E-04 | 5.3153E-02 | 3.3474E-04 | -5.6928E+00 | 3.7799E+01 |
| | 测试集 | 7.8396E-02 | 6.7692E-04 | 6.0922E-02 | 4.1785E-04 | -8.7153E+00 | 4.7630E+01 |
| **BHFR** | 训练集 | 6.0665E-03 | 1.6330E-08 | 3.9665E-03 | 8.4708E-09 | 9.5669E-01 | 3.3481E-06 |
| | 验证集 | 2.1551E-02 | 3.5181E-08 | **1.2384E-02** | 3.5083E-08 | 4.2484E-01 | 9.8731E-05 |
| | 测试集 | **1.6189E-02** | 2.2474E-08 | **1.1226E-02** | **1.0102E-08** | **6.2491E-01** | 4.8607E-05 |

[0271] 由表1和图4a-4c可知:1)RF在训练、验证和测试中的RMSE、MAE和$R^2$指标均值统计结果均优于DFR,但在稳定性指标上弱于DFR;2)DFR和DFR-clfc,在建模精度上与RF接近,同时建模稳定性要好于RF,其中DFR-clfc在训练、验证和测试集的精度略高于DFR,但DFR的稳定性更好;3)BLS-NN对训练数据出现了明显的过拟合,其在验证和测试集中的泛化性能和稳定性上均表现最差,表明BLS-NN难以适用于本申请中的真实工业过程的小样本高维数据;4)BHFR在测试集中的RMSE、MAE和$R^2$指标的均值统计结果均为最佳,稳定性仅弱于DFR,表明BHFR具有良好的泛化性能和稳定性。

[0272] 综上可知,DXN软测量建模实验表明本申请所提BHFR具有比经典RF、DFR极其改进版DFR-clfc更好的训练学习能力,同时在测试集上的建模精度和对数据的拟合程度也强于RF、DFR、DFR-clfc和BLS-NN,体现了其在构建DXN软测量模型中的明显优势。

[0273] 本发明提供的基于宽度混合森林回归的MSWI过程二噁英排放软测量方法,建立了基于BHFR的软测量模型,其结合了宽度学习建模、集成学习和潜在特征提取等算法,1)基于宽度学习系统框架,采用非微分学习器构建了包含特征映射层、潜在特征提取层、特征增强层和增量学习层的软测量模型;2)利用信息全联接、潜在特征提取和互信息度量对BHFR模型内部信息进行处理,有效保证了BHFR模型内部特征信息的传递最大化和冗余度最小化;3)采用混合森林组为映射单元实现建模过程的增量学习,通过伪逆策略快速计算输出层权重矩阵,再利用训练误差的收敛程度自适应调整增量学习,实现了高精度的软测量建模。在高维基准数据集和工业过程DXN数据集上验证了所提方法的有效性和合理性。

[0274] 本文中应用了具体个例对本发明的原理及实施方式进行了阐述,以上实施例的说明只是用于帮助理解本发明的方法及其核心思想;同时,对于本领域的一般技术人员,依据本发明的思想,在具体实施方式及应用范围上均会有改变之处。综上所述,本说明书内容不应理解为对本发明的限制。

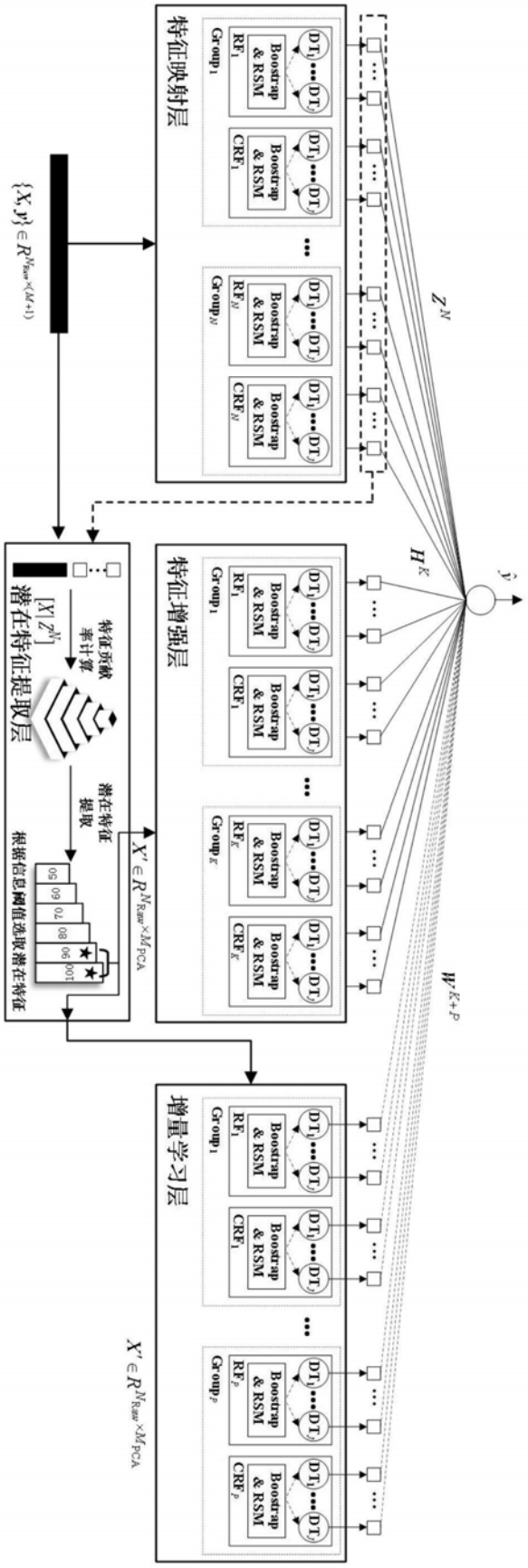

特征映射层
特征增强层
潜在特征提取层
增量学习层
特征贡献
潜在特征
提取
根据信息阈值选取潜在特征
Boostrap & RSM
Group
RF
CRF
DT

图1

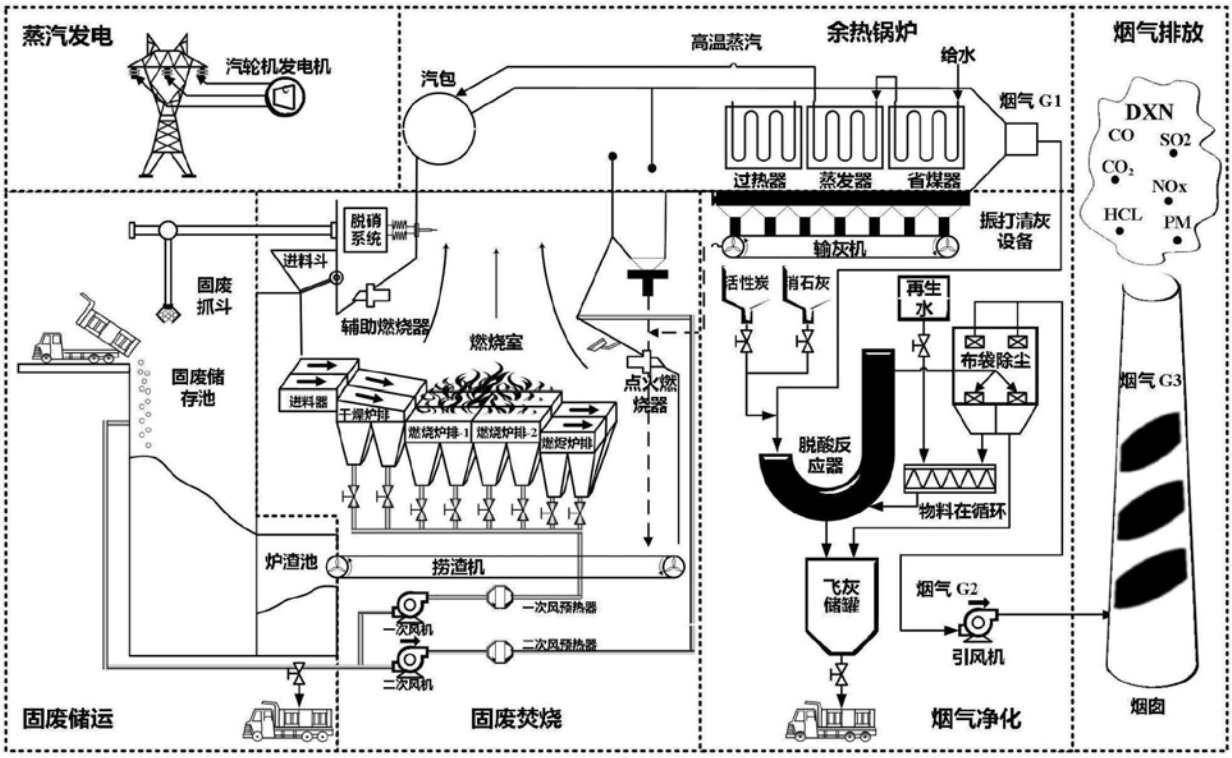

蒸汽发电
汽轮机发电机
余热锅炉
高温蒸汽
汽包
给水
烟气 G1
过热器
蒸发器
省煤器
烟气排放
DXN
CO
SO2
CO2
NOx
HCL
PM
脱硝系统
进料斗
固废抓斗
辅助燃烧器
燃烧室
点火燃烧器
振打清灰设备
输灰机
活性炭
消石灰
再生水
布袋除尘
烟气 G3
固废储存池
进料器
干燥炉排
燃烧炉排-1
燃烧炉排-2
燃烬炉排
脱酸反应器
物料在循环
炉渣池
捞渣机
一次风预热器
一次风机
二次风预热器
二次风机
飞灰储罐
烟气 G2
引风机
烟囱
固废储运
固废焚烧
烟气净化

图2

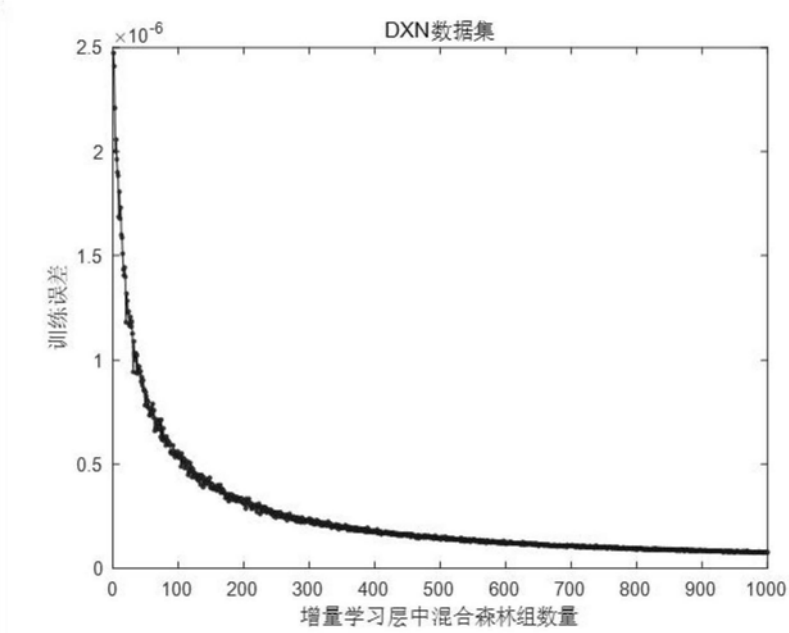

DXN数据集
×10-6
2.5
2
1.5
1
0.5
0
训练误差
0
100
200
300
400
500
600
700
800
900
1000
增量学习层中混合森林组数量

图3

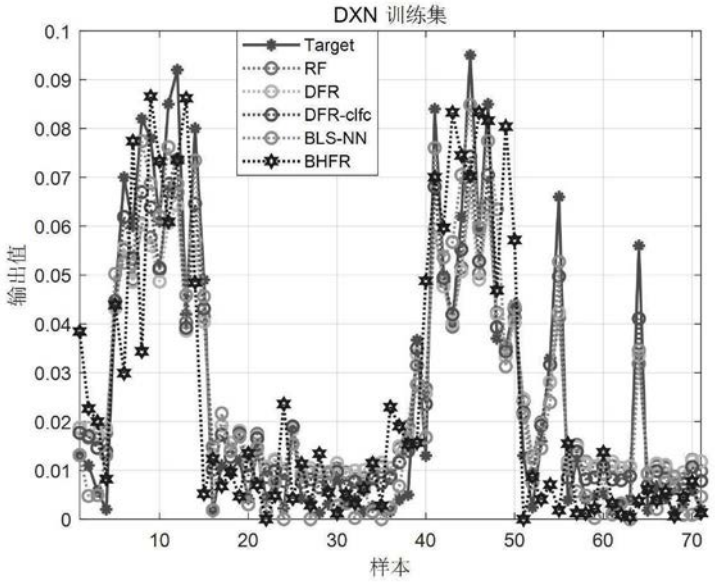

DXN 训练集
Target
RF
DFR
DFR-clfc
BLS-NN
BHFR
输出值
样本
0.1
0.09
0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0
10
20
30
40
50
60
70

图4a

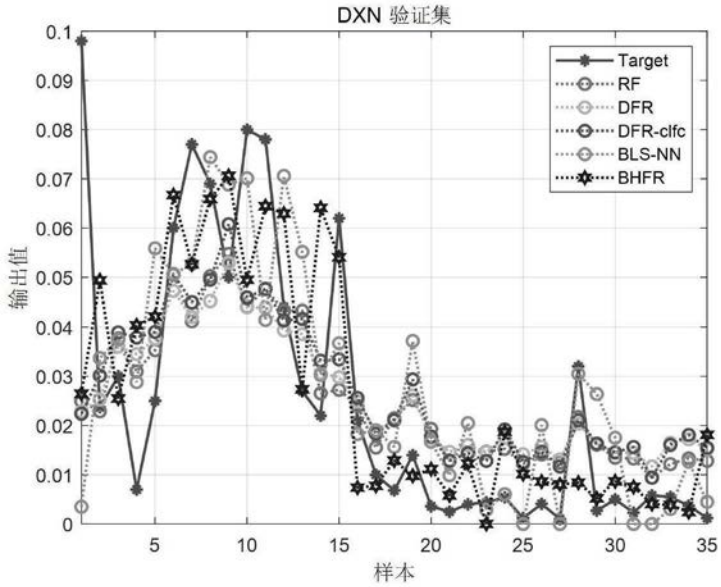

DXN 验证集
Target
RF
DFR
DFR-clfc
BLS-NN
BHFR
输出值
样本
0.1
0.09
0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0
5
10
15
20
25
30
35

图4b

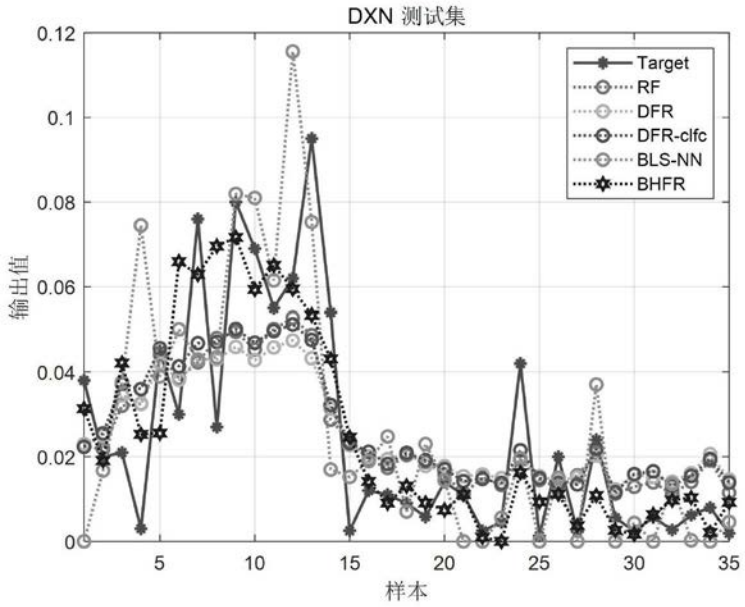

DXN 测试集
Target
RF
DFR
DFR-clfc
BLS-NN
BHFR
输出值
样本
0.12
0.1
0.08
0.06
0.04
0.02
0
5
10
15
20
25
30
35

图4c