(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 114266461 A
(43)申请公布日 2022.04.01

(21)申请号 202111539001.X

(22)申请日 2021.12.15

(71)申请人 北京工业大学
地址 100124 北京市朝阳区平乐园100号

(72)发明人 汤健 崔璨麟 夏恒 王丹丹 乔俊飞

(74)专利代理机构 北京思海天达知识产权代理有限公司 11203
代理人 刘萍

(51)Int.Cl.
G06Q 10/06(2012.01)
G06Q 50/26(2012.01)
G06K 9/62(2022.01)

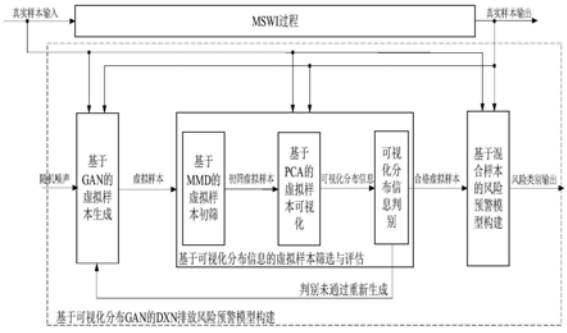
权利要求书3页 说明书9页 附图5页

(54)发明名称

基于可视化分布GAN的MSWI过程二噁英排放风险预警方法

(57)摘要

基于可视化分布GAN的MSWI过程二噁英排放风险预警方法属于城市固废焚烧领域。目前工业过程中采用长周期、高成本的离线剧毒污染物二噁英DXN排放浓度检测方式，这使得用于构建风险预警模型的样本极其稀少。针对上述问题，提出基于可视化分布生成对抗网络(GAN)的MSWI过程DXN排放风险预警建模方法。首先，在原始GAN的基础上引入DXN的风险等级作为条件信息，使得生成器可以生成指定风险等级的虚拟样本。然后，利用可视化分布信息评估并筛选合格虚拟样本。最后，基于虚拟样本和真实样本组成的混合样本构建DXN排放风险预警模型。通过工业过程DXN数据验证了所提方法的有效性。



CN 114266461 A

1.基于可视化分布GAN的MSWI过程二噁英排放风险预警方法,其特征在于:

真实样本输入和对应的输出分别记为X_{real}和Y_{real};随机噪声记为X_{noise};GAN生成器生成的虚拟样本记为$(\boldsymbol{X}_{vir}^{ori},\boldsymbol{Y}_{vir}^{ori})$,其中$\boldsymbol{X}_{vir}^{ori}$表示虚拟样本输入集,$\boldsymbol{Y}_{vir}^{ori}$表示对应的虚拟样本输出集;经过MMD初筛的虚拟样本记为$(\boldsymbol{X}_{vir}^{inisel},\boldsymbol{Y}_{vir}^{inisel})$,其中$\boldsymbol{X}_{vir}^{inisel}$表示初筛虚拟样本输入集,$\boldsymbol{Y}_{vir}^{inisel}$表示对应初筛虚拟样本输出集;可视化分布信息记为$D_{PCA}$;经过可视化分布信息判别得到的合格虚拟样本记为$(\boldsymbol{X}_{vir}^{fine},\boldsymbol{Y}_{vir}^{fine})$,其中$\boldsymbol{X}_{vir}^{fine}$表示合格虚拟样本输入集,$\boldsymbol{Y}_{vir}^{fine}$表示对应的合格虚拟样本输出集;所构建风险预警模型的风险类别输出记为$\hat{\boldsymbol{Y}}$;

1)基于GAN的虚拟样本生成模块

虚拟样本生成的流程为:首先,将X_{noise}和Y_{real}共同输入生成器以生成虚拟样本的输入$\boldsymbol{X}_{vir}^{train}$;接着,将$X_{real}$、$\boldsymbol{X}_{vir}^{train}$和$Y_{real}$再输入判别器,根据判别结果$Y_{real/vir}$更新生成器和判别器;然后,将$X_{noise}$和期望生成的DXN排放风险等级$\boldsymbol{Y}_{vir}^{ori}$输入训练好的生成器以生成$\boldsymbol{X}_{vir}^{ori}$;最后,将$\boldsymbol{X}_{vir}^{ori}$和$\boldsymbol{Y}_{vir}^{ori}$组合,得到虚拟样本$(\boldsymbol{X}_{vir}^{ori},\boldsymbol{Y}_{vir}^{ori})$;

每批训练样本数设为N_b,学习率为α_{Ir},最大训练代数为N_e;生成器采用三层神经网络,隐含层使用Relu激活函数,输出层使用线性激活函数,如下:

$$\begin{cases}\boldsymbol{H}_{G}^{hidden}=\text{relu}((\boldsymbol{X}_{noise},\boldsymbol{Y}_{real})\cdot\boldsymbol{\omega}_{G1}+\boldsymbol{b}_{G1})\\ \boldsymbol{X}_{vir}^{train}=\boldsymbol{H}_{G}^{hidden}\cdot\boldsymbol{\omega}_{G2}+\boldsymbol{b}_{G2}\end{cases}\tag{1}$$

其中,ω_{G1}为生成器输入层和隐含层之间的权值;b_{G1}为生成器输入层和隐含层之间的偏置;Relu激活函数relu(x)=max(0,x),x为任意输入值;$\boldsymbol{H}_{G}^{hidden}$为生成器隐含层输出;$\omega_{G2}$为生成器隐含层和输出层之间的权值;$b_{G2}$为生成器隐含层和输出层之间的偏置;

判别器采用三层神经网络,隐含层使用Relu激活函数,输出层使用Sigmoid激活函数,如下:

$$\begin{cases}\boldsymbol{S}_{mix}^{train}=\{(\boldsymbol{X}_{vir}^{train},\boldsymbol{Y}_{real}),(\boldsymbol{X}_{real},\boldsymbol{Y}_{real})\}\\ \boldsymbol{H}_{D}^{hidden}=\text{relu}(\boldsymbol{S}_{mix}^{train}\cdot\boldsymbol{\omega}_{D1}+\boldsymbol{b}_{D1})\\ \boldsymbol{Y}_{real/vir}=\text{sigmoid}(\boldsymbol{H}_{G}^{hidden}\cdot\boldsymbol{\omega}_{D2}+\boldsymbol{b}_{D2})\end{cases}\tag{2}$$

其中,$\boldsymbol{S}_{mix}^{train}$为$(\boldsymbol{X}_{vir}^{train},\boldsymbol{Y}_{real})$和$(X_{real},Y_{real})$组成的混合样本;$\omega_{D1}$为判别器输入层和隐含层之间的权值;$b_{D1}$为判别器输入层和隐含层之间的偏置;$\boldsymbol{H}_{D}^{hidden}$为判别器隐含层输出;$\omega_{D2}$为判别器隐含层和输出层之间的权值;$b_{D2}$为判别器隐含层和输出层之间的偏置;Sigmoid激活函数$\text{sigmoid}(x)=\dfrac{1}{1+e^{-x}}$,x为任意输入值;

目标函数O_{GAN}如式(3)所示:

$$O_{GAN}=E_{P_{data}(\boldsymbol{X}_{real})}(\log\boldsymbol{Y}_{D}^{real})+E_{P_{noise}(\boldsymbol{X}_{noise})}[\log(1-\boldsymbol{Y}_{D}^{vir})]\tag{3}$$

其中,$P_{data}(X_{real})$表示X_{real}的分布;$\boldsymbol{Y}_{D}^{real}$为判别器对于$(X_{real},Y_{real})$的输出;$P_{noise}(X_{noise})$表示$X_{noise}$的分布;$\boldsymbol{Y}_{D}^{vir}$为判别器对于$(\boldsymbol{X}_{vir}^{train},\boldsymbol{Y}_{real})$的输出;

判别器计算样本是来自$P_{noise}(X_{noise})$还是$P_{data}(X_{real})$的概率，生成器根据判别器结果学习真实样本的分布$P_{data}(X_{real})$以减少$\log(1-\boldsymbol{Y}_{\mathrm{D}}^{\mathrm{vir}})$，生成器和判别器在最小最大的博弈对抗中共同训练；

2)基于可视化分布信息的虚拟样本筛选与评估模块

基于MMD的虚拟样本初筛模块

首先，取若干个生成器，生成若干组虚拟样本；

接着，计算每组虚拟样本质量；采用MMD度量虚拟样本与真实样本的总体均值差异，进而衡量两者之间的分布差异；

假设，$\boldsymbol{X}_{\mathrm{vir}}^{\mathrm{ori}}=\{x_1^{(s)},x_2^{(s)},\cdots,x_{N_{\mathrm{vir}}^{\mathrm{ori}}}^{(s)}\}$服从分布$P_{\mathrm{vir}}^{\mathrm{ori}}$，其中$N_{\mathrm{vir}}^{\mathrm{ori}}$是一组虚拟样本数量；$\boldsymbol{X}_{\mathrm{real}}=\{x_1^{(t)},x_2^{(t)},\cdots,x_{N_{\mathrm{real}}}^{(t)}\}$服从分布$P_{real}$，$N_{real}$是真实样本数量；通过高维映射函数获得两个域样本在再生希尔伯特空间中期望差值的上确界，即：

$$\begin{aligned}&\mathrm{MMD}(\boldsymbol{X}_{\mathrm{vir}}^{\mathrm{ori}},\boldsymbol{X}_{\mathrm{real}})=\sup_{\phi\in H}\|E_p[\phi(\boldsymbol{X}_{\mathrm{vir}}^{\mathrm{ori}})]-E_q[\phi(\boldsymbol{X}_{\mathrm{real}})]\|_H\\&=\|\frac{1}{N_{\mathrm{vir}}^{\mathrm{ori}}}\sum_{i=1}^{N_{\mathrm{vir}}^{\mathrm{ori}}}\phi(x_i^{(s)})-\frac{1}{N_{\mathrm{real}}}\sum_{j=1}^{N_{\mathrm{real}}}\phi(x_j^{(t)})\|_H\\&=\frac{1}{(N_{\mathrm{vir}}^{\mathrm{ori}})^2}\sum_{i,j=1}^{N_{\mathrm{vir}}^{\mathrm{ori}}}\exp(\frac{-\|x_i^{(s)}-x_j^{(s)}\|}{\sigma})+\frac{1}{(N_{\mathrm{real}})^2}\sum_{i,j=1}^{N_{\mathrm{real}}}\exp(\frac{-\|x_i^{(t)}-x_j^{(t)}\|}{\sigma})\\&-\frac{2}{N_{\mathrm{vir}}^{\mathrm{ori}}N_{\mathrm{real}}}\sum_{i,j=1}^{N_{\mathrm{vir}}^{\mathrm{ori}},N_{\mathrm{real}}}\exp(\frac{-\|x_i^{(s)}-x_j^{(t)}\|}{\sigma})\end{aligned}\tag{4}$$

其中，H为RKHS，φ(•)表示将样本映射到高维RKHS，$E_p[\phi(\boldsymbol{X}_{\mathrm{vir}}^{\mathrm{ori}})]$和$E_q[\phi(X_{real})]$表示样本映射到RKHS中的期望值，σ是高斯核的带宽；

根据式(4)分别计算N组虚拟样本$\{(\boldsymbol{X}_1^{\mathrm{vir}},\boldsymbol{Y}_1^{\mathrm{vir}}),(\boldsymbol{X}_2^{\mathrm{vir}},\boldsymbol{Y}_2^{\mathrm{vir}}),\cdots,(\boldsymbol{X}_N^{\mathrm{vir}},\boldsymbol{Y}_N^{\mathrm{vir}})\}$与真实样本$(X_{real},Y_{real})$的MMD值以对其进行初筛，$(\boldsymbol{X}_1^{\mathrm{vir}},\boldsymbol{Y}_1^{\mathrm{vir}})$为第一组虚拟样本，$(\boldsymbol{X}_2^{\mathrm{vir}},\boldsymbol{Y}_2^{\mathrm{vir}})$为第二组虚拟样本，$(\boldsymbol{X}_N^{\mathrm{vir}},\boldsymbol{Y}_N^{\mathrm{vir}})$为第N组虚拟样本，初筛函数如下：

$$\begin{aligned}&\phi_{\mathrm{MMD}}((\boldsymbol{X}_1^{\mathrm{vir}},\boldsymbol{Y}_1^{\mathrm{vir}}),(\boldsymbol{X}_2^{\mathrm{vir}},\boldsymbol{Y}_2^{\mathrm{vir}}),\cdots,(\boldsymbol{X}_N^{\mathrm{vir}},\boldsymbol{Y}_N^{\mathrm{vir}}))=\\&\min(\mathrm{MMD}(\boldsymbol{X}_1^{\mathrm{vir}},\boldsymbol{X}_{\mathrm{real}}),\mathrm{MMD}(\boldsymbol{X}_2^{\mathrm{vir}},\boldsymbol{X}_{\mathrm{real}}),\cdots,\mathrm{MMD}(\boldsymbol{X}_N^{\mathrm{vir}},\boldsymbol{X}_{\mathrm{real}}))\end{aligned}\tag{5}$$

其中，min(•)表示取N组虚拟样本与(X_{real},Y_{real})MMD值最小的那组虚拟样本，即其作为质量最好的初筛虚拟样本$(\boldsymbol{X}_{\mathrm{vir}}^{\mathrm{inisel}},\boldsymbol{Y}_{\mathrm{vir}}^{\mathrm{inisel}})$；

基于PCA的虚拟样本可视化模块

采用PCA将虚拟DXN样本降到1维进行可视化以提供整体分布信息

PCA通过一组正交向量将原始数据投影到新的空间，消除了原始数据冗余的同时保留了主要信息；基于PCA的虚拟样本可视化实现步骤如下；

首先，对$\boldsymbol{X}_{\mathrm{vir}}^{\mathrm{inisel}}$进行中心化处理，得到中心化样本U，其中$N_{\mathrm{vir}}^{\mathrm{inisel}}$为样本数量，$M_{\mathrm{vir}}^{\mathrm{inisel}}$为样本维数；

接着，计算U的协方差矩阵C：

$C=UU^T$ (6)

然后,采用特征分解法计算C的特征向量和特征值:

$C=W\Lambda W^T$ (7)

其中,W为特征向量组成的矩阵;Λ为特征根按照从大到小顺序排列的对角阵;

再然后,将$\boldsymbol{X}_{vir}^{inisel}$降到1维,如下:

$X_{PCA}=\mu_1 U$ (8)

式中,X_{PCA}为降到1维的虚拟样本;μ_1为最大特征值对应的特征向量;

最后,计算X_{PCA}分布函数,得到PCA可视化结果;

可视化分布信息判别模块

虚拟样本PCA可视化结果提供的整体分布信息D_{PCA}:

$D_{PCA}=(R_{real}\cap R_{vir})/S_{real}$ (9)

其中,R_{real}为真实样本分布函数与x轴包含的区域;R_{vir}为虚拟样本分布函数与x轴包含的区域;S_{real}为真实样本分布函数与x轴包含的区域的面积;$R_{real}\cap R_{vir}$表示R_{real}与R_{vir}重叠部分的面积;

可视化分布信息判别函数如下:

$$\phi_{visual}(D_{PCA})=\begin{cases}1\ ,\ if\ D_{PCA}\geq\theta_S\\0\ ,\ else\ D_{PCA}<\theta_S\end{cases} \quad (10)$$

其中,θ_S为设定的阈值,取0.8,

若$\Phi_{visual}(D_{PCA})$的值为1,表示该虚拟样本为合格虚拟样本;反之,该虚拟样本为不合格虚拟样本;

3)基于混合样本的风险预警模型构建模块

判别得到的合格虚拟样本$(\boldsymbol{X}_{vir}^{fine},\boldsymbol{Y}_{vir}^{fine})$和真实样本$(X_{real},Y_{real})$组合,得到混合样本$S_{mix}$;

$$\boldsymbol{S}_{mix}=\{(\boldsymbol{X}_{vir}^{fine},\boldsymbol{Y}_{vir}^{fine}),(\boldsymbol{X}_{real},\boldsymbol{Y}_{real})\} \quad (11)$$

使用随机森林作为风险预警模型的分类器,步骤如下;

首先,利用Bootstrap算法和RSM算法对S_{mix}进行样本和特征的随机采样,获得N个子样本集;

接着,利用N个子样本集构建N个决策树,每个决策树得到一个分类结果;

最后,对N个分类结果进行投票,选择投票数量最多的类别作为最终的分类结果$\hat{\boldsymbol{Y}}$。

基于可视化分布GAN的MSWI过程二噁英排放风险预警方法

技术领域

[0001] 本发明属于城市固废焚烧领域。

背景技术

[0002] 城市固废(Municipal Solid Waste,MSW)的产生量随城市人口的不断增加而逐年提高。城市固废焚烧(MSW Incineration,MSWI)是当今世界大部分国家采用的具有无害化、减量化和资源化等优势的处理手段。由于MSWI过程所产生的副产品二噁英(Dioxins,DXN)为剧毒污染物,不但损害中毒者的内分泌系统和破坏染色体进而导致细胞癌变,而且在生物体内具有累积效应,是造成焚烧建厂存在“邻避效应”的主要原因。因此,控制其排放是亟需解决的环保问题。对DXN排放的风险等级进行预警,进而优化控制MSWI过程,对减少污染物排放具有重要的实际意义。

[0003] 目前,工业界主要对MSWI过程末端烟囱排放烟气中的DXN进行检测。离线直接检测法和在线间接检测法均很难满足MSWI过程以减少DXN排放为目的的实时优化控制。此外,由于DXN排放浓度检测的难度大、周期长、费用昂贵,导致构建数据驱动模型的样本真值极其稀少。因此,本申请所研究的MSWI过程DXN排放浓度检测问题属于典型的小样本问题,具有样本数量少、样本不平衡等特性。通常而言,较少数量的建模样本难以准确反映工业过程的真实特性,难以构建鲁棒可靠的污染物浓度排放回归预测模型;相对而言,构建风险判别分类模型较为容易。此外,工业现场领域专家也常习惯于用排放浓度的低、中、高等等级语言描述污染排放程度的风险,并依据自身经验获得判别结果以调整相关控制参数。但是,样本的不平衡,即某类样本的数量远小于其他类,这也是导致所构建的风险判别模型具有片面性和偏差性的主要原因。

[0004] 综上,本申请提出基于主动学习机制GAN的MSWI过程DXN排放风险预警模型构建方法。首先,在原始GAN的基础上引入DXN风险等级作为条件信息,将其与随机噪声输入生成器后生成预设定DXN风险等级的虚拟样本,与真实样本共同输入判别器后根据判别结果更新生成器和判别器;接着,先使用最大均值差异(Maximum Mean Discrepancy,MMD)对虚拟样本进行初筛,再对初筛虚拟样本采用主成分分析(Principal Component Analysis,PCA)以获得可视化分布信息,根据其判断初筛虚拟样本是否合格;最后,基于虚拟样本和真实样本组成的混合样本构建DXN排放风险预警模型。结合实际DXN数据验证了所提方法的有效性。

[0005] 国内某MSWI电厂的炉排炉焚烧工艺流程如图1所示。

[0006] 由图1可知,MSW由专用车辆收集、称重后运输至卸料大厅,倾倒入密封的固废池中,并通过抓斗送至焚烧炉料斗内,由给料器推至炉排;MSW在焚烧炉内依次经历干燥、点燃、燃烧和燃烬四个阶段,燃烬后的残渣落入水冷渣斗后由捞渣机送至灰渣坑中,收集后送至填埋场处理;焚烧过程产生的烟气加热余热锅炉产生高压蒸汽进而推动汽轮发电机发电;添加活性炭和消石灰后,锅炉出口烟气中进入反应器,产生的飞灰进入飞灰储罐,烟气进入袋式除尘器去除烟气颗粒物、中和反应物和活性炭吸附物,处理之后分为三部分:尾部飞灰进入飞灰罐,部分烟灰混合物在混合器中加水后重新进入反应器,尾部烟气通过引风

机经烟囱排入大气，其中包含HCL、SO_2、CO、CO_2、NO_x和DXN等物质。

[0007] 由于固废不完全燃烧和新规合成反应生成两种原因导致MSWI过程产生的焚烧灰、飞灰和烟气中包含DXN。因此，焚烧过程中烟气需要达到850℃并保持2s以确保有毒有机物的有效分解。在烟气处理阶段向反应器内注入石灰和活性炭，吸附DXN和部分重金属，然后经袋式除尘器过滤通过引风机排入烟囱，以减少排放烟气中的DXN浓度。此外，该阶段产生的积灰存在的DXN记忆效应也会导致DXN排放浓度增加。现场分布式控制系统(DCS)采集和存储上述各阶段与DXN相关的过程变量以及常规污染物(CO、HCL、SO_2、NO_x和HF等)浓度。然而，由于高成本和长周期等原因使得排放烟气中DXN的检测较为困难。

[0008] 由上可知，构建DXN排放风险预警模型的样本存在数量少、分布不均和维数高等特点。因此，本申请提出一种基于主动学习机制虚拟样本对抗生成策略建立MSWI过程DXN排放风险预警建模。

发明内容

[0009] 本申请提出的基于主动学习机制GAN的MSWI过程DXN排放风险预警模型构建策略，包括：基于GAN的虚拟样本生成、基于可视化分布信息的虚拟样本评估与筛选和基于混合样本的风险预警模型构建三个模块，如图2所示。

[0010] 在图2中，真实样本输入和对应的输出分别记为X_{real}和Y_{real}；随机噪声记为X_{noise}；GAN生成器生成的虚拟样本记为$(\boldsymbol{X}_{vir}^{ori},\boldsymbol{Y}_{vir}^{ori})$，其中$\boldsymbol{X}_{vir}^{ori}$表示虚拟样本输入集，$\boldsymbol{Y}_{vir}^{ori}$表示对应的虚拟样本输出集；经过MMD初筛的虚拟样本记为$(\boldsymbol{X}_{vir}^{inisel},\boldsymbol{Y}_{vir}^{inisel})$，其中$\boldsymbol{X}_{vir}^{inisel}$表示初筛虚拟样本输入集，$\boldsymbol{Y}_{vir}^{inisel}$表示对应初筛虚拟样本输出集；可视化分布信息记为$D_{PCA}$；经过可视化分布信息判别得到的合格虚拟样本记为$(\boldsymbol{X}_{vir}^{fine},\boldsymbol{Y}_{vir}^{fine})$，其中$\boldsymbol{X}_{vir}^{fine}$表示合格虚拟样本输入集，$\boldsymbol{Y}_{vir}^{fine}$表示对应的合格虚拟样本输出集；所构建风险预警模型的风险类别输出记为$\hat{\boldsymbol{Y}}$。

[0011] 该策略不同模块的功能如下：

[0012] 1)基于GAN的虚拟样本生成模块：在原始GAN的基础上引入DXN排放风险等级作为条件信息，将其和随机噪声共同输入生成器以生成指定类型的虚拟样本；进一步，将虚拟样本和真实样本再输入判别器，并根据判别结果更新生成器和判别器；最后，在生成器和判别器的博弈对抗中，使得生成的虚拟样本越来越接近真实样本。

[0013] 2)基于可视化分布信息的虚拟样本筛选与评估模块：首先，使用MMD计算虚拟样本和真实样本的相似程度对虚拟样本进行初筛；然后，基于PCA进行虚拟样本可视化以获得降维后的分布信息；最后，依据分布信息进行判别并确定是否合格，若是，则标定为合格虚拟样本；若否，则重新生成虚拟样本。

[0014] 3)基于混合样本的风险预警模型构建模块：基于混合样本采用随机森林算法构建DXN排放风险预警模型。

[0015] 4.1基于GAN的虚拟样本生成模块

[0016] GAN是一种基于博弈场景的无监督生成模型，通过生成器和判别器的博弈对抗生成接近真实样本的虚拟样本。由于原始GAN生成的虚拟样本类型不可控，本模块在原始GAN的基础上引入DXN排放风险等级作为条件信息控制生成虚拟样本的类型。

[0017] 基于GAN的虚拟样本生成流程如图3所示。

[0018] 虚拟样本生成的流程为:首先,将X_{noise}和Y_{real}共同输入生成器以生成虚拟样本的输入$\boldsymbol{X}_{vir}^{train}$;接着,将$X_{real}$、$\boldsymbol{X}_{vir}^{train}$和$Y_{real}$再输入判别器,根据判别结果$Y_{real/vir}$更新生成器和判别器;然后,将$X_{noise}$和期望生成的DXN排放风险等级$\boldsymbol{Y}_{vir}^{ori}$输入训练好的生成器以生成$\boldsymbol{X}_{vir}^{ori}$;最后,将$\boldsymbol{X}_{vir}^{ori}$和$\boldsymbol{Y}_{vir}^{ori}$组合,得到虚拟样本$(\boldsymbol{X}_{vir}^{ori},\boldsymbol{Y}_{vir}^{ori})$。

[0019] 本申请中,每批训练样本数设为N_b,学习率为α_{Ir},最大训练代数为N_e。生成器采用三层神经网络,隐含层使用Relu激活函数,输出层使用线性激活函数,如下:

[0020]
$$\begin{cases} \boldsymbol{H}_{G}^{hidden}=\mathrm{relu}((\boldsymbol{X}_{noise},\boldsymbol{Y}_{real})\cdot\boldsymbol{\omega}_{G1}+\boldsymbol{b}_{G1}) \\ \boldsymbol{X}_{vir}^{train}=\boldsymbol{H}_{G}^{hidden}\cdot\boldsymbol{\omega}_{G2}+\boldsymbol{b}_{G2} \end{cases} \tag{1}$$

[0021] 其中,ω_{G1}为生成器输入层和隐含层之间的权值;b_{G1}为生成器输入层和隐含层之间的偏置;Relu激活函数relu(x)=max(0,x),x为任意输入值;$\boldsymbol{H}_{G}^{hidden}$为生成器隐含层输出;$\omega_{G2}$为生成器隐含层和输出层之间的权值;$b_{G2}$为生成器隐含层和输出层之间的偏置。

[0022] 判别器采用三层神经网络,隐含层使用Relu激活函数,输出层使用Sigmoid激活函数,如下:

[0023]
$$\begin{cases} \boldsymbol{S}_{mix}^{train}=\{(\boldsymbol{X}_{vir}^{train},\boldsymbol{Y}_{real}),(\boldsymbol{X}_{real},\boldsymbol{Y}_{real})\} \\ \boldsymbol{H}_{D}^{hidden}=\mathrm{relu}(\boldsymbol{S}_{mix}^{train}\cdot\boldsymbol{\omega}_{D1}+\boldsymbol{b}_{D1}) \\ \boldsymbol{Y}_{real/vir}=\mathrm{sigmoid}(\boldsymbol{H}_{G}^{hidden}\cdot\boldsymbol{\omega}_{D2}+\boldsymbol{b}_{D2}) \end{cases} \tag{2}$$

[0024] 其中,$\boldsymbol{S}_{mix}^{train}$为$(\boldsymbol{X}_{vir}^{train},\boldsymbol{Y}_{real})$和$(X_{real},Y_{real})$组成的混合样本;$\omega_{D1}$为判别器输入层和隐含层之间的权值;$b_{D1}$为判别器输入层和隐含层之间的偏置;$\boldsymbol{H}_{D}^{hidden}$为判别器隐含层输出;$\omega_{D2}$为判别器隐含层和输出层之间的权值;$b_{D2}$为判别器隐含层和输出层之间的偏置;Sigmoid激活函数$\mathrm{sigmoid}(x)=\frac{1}{1+e^{-x}}$,x为任意输入值。

[0025] 目标函数O_{GAN}如式(3)所示:

[0026]
$$O_{GAN}=E_{P_{data}(\boldsymbol{X}_{real})}(\log\boldsymbol{Y}_{D}^{real})+E_{P_{noise}(\boldsymbol{X}_{noise})}[\log(1-\boldsymbol{Y}_{D}^{vir})] \tag{3}$$

[0027] 其中,$P_{data}(X_{real})$表示X_{real}的分布;$\boldsymbol{Y}_{D}^{real}$为判别器对于$(X_{real},Y_{real})$的输出;$P_{noise}(X_{noise})$表示$X_{noise}$的分布;$\boldsymbol{Y}_{D}^{vir}$为判别器对于$(\boldsymbol{X}_{vir}^{train},\boldsymbol{Y}_{real})$的输出。

[0028] 判别器计算样本是来自$P_{noise}(X_{noise})$还是$P_{data}(X_{real})$的概率,生成器根据判别器结果学习真实样本的分布$P_{data}(X_{real})$以减少$\log(1-Y_{D}^{vir})$,生成器和判别器在最小最大的博弈对抗中共同训练。

[0029] 4.2基于可视化分布信息的虚拟样本筛选与评估模块

[0030] 基于可视化分布信息的虚拟样本筛选与评估流程是:首先,根据虚拟样本和真实样本的MMD值对虚拟样本进行初筛;然后,对虚拟样本的PCA可视化分布信息进行最终判断;最后,若判别不通过则重新生成虚拟样本,继续执行上述操作,若判别通过则得到合格的虚拟样本。流程如图4所示。

[0031] 4.2.1基于MMD的虚拟样本初筛模块

[0032] 首先,取若干个生成器,生成若干组虚拟样本。

[0033] 接着,计算每组虚拟样本质量。本申请中,采用MMD度量虚拟样本与真实样本的总体均值差异,进而衡量两者之间的分布差异。

[0034] 假设,$\boldsymbol{X}_{\text{vir}}^{\text{ori}}=\{x_1^{(s)},x_2^{(s)},\cdots,x_{N_{\text{vir}}^{\text{ori}}}^{(s)}\}$服从分布$P_{\text{vir}}^{\text{ori}}$,其中$N_{\text{vir}}^{\text{ori}}$是一组虚拟样本数量;$\boldsymbol{X}_{\text{real}}=\{x_1^{(t)},x_2^{(t)},\cdots,x_{N_{\text{real}}}^{(t)}\}$服从分布$P_{\text{real}}$,$N_{\text{real}}$是真实样本数量。进一步,通过高维映射函数获得两个域样本在再生希尔伯特空间(Reproducing kernel Hilbert space,RKHS)中期望差值的上确界,即:

[0035]
$$\begin{aligned}&\text{MMD}(\boldsymbol{X}_{\text{vir}}^{\text{ori}},\boldsymbol{X}_{\text{real}})=\sup_{\phi\in H}\|E_p[\phi(\boldsymbol{X}_{\text{vir}}^{\text{ori}})]-E_q[\phi(\boldsymbol{X}_{\text{real}})]\|_H\\&=\|\frac{1}{N_{\text{vir}}^{\text{ori}}}\sum_{i=1}^{N_{\text{vir}}^{\text{ori}}}\phi(x_i^{(s)})-\frac{1}{N_{\text{real}}}\sum_{j=1}^{N_{\text{real}}}\phi(x_j^{(t)})\|_H\\&=\frac{1}{(N_{\text{vir}}^{\text{ori}})^2}\sum_{i,j=1}^{N_{\text{vir}}^{\text{ori}}}\exp(\frac{-\|x_i^{(s)}-x_j^{(s)}\|}{\sigma})+\frac{1}{(N_{\text{real}})^2}\sum_{i,j=1}^{N_{\text{real}}}\exp(\frac{-\|x_i^{(t)}-x_j^{(t)}\|}{\sigma})\\&-\frac{2}{N_{\text{vir}}^{\text{ori}}N_{\text{real}}}\sum_{i,j=1}^{N_{\text{vir}}^{\text{ori}},N_{\text{real}}}\exp(\frac{-\|x_i^{(s)}-x_j^{(t)}\|}{\sigma})\end{aligned}\tag{4}$$

[0036] 其中,H为RKHS,φ(•)表示将样本映射到高维RKHS,$E_p[\phi(\boldsymbol{X}_{\text{vir}}^{\text{ori}})]$和$E_q[\phi(X_{\text{real}})]$表示样本映射到RKHS中的期望值,σ是高斯核的带宽。

[0037] 根据式(4)分别计算N组虚拟样本$\{(\boldsymbol{X}_1^{\text{vir}},\boldsymbol{Y}_1^{\text{vir}}),(\boldsymbol{X}_2^{\text{vir}},\boldsymbol{Y}_2^{\text{vir}}),\cdots,(\boldsymbol{X}_N^{\text{vir}},\boldsymbol{Y}_N^{\text{vir}})\}$与真实样本$(X_{\text{real}},Y_{\text{real}})$的MMD值以对其进行初筛,$(\boldsymbol{X}_1^{\text{vir}},\boldsymbol{Y}_1^{\text{vir}})$为第一组虚拟样本,$(\boldsymbol{X}_2^{\text{vir}},\boldsymbol{Y}_2^{\text{vir}})$为第二组虚拟样本,$(\boldsymbol{X}_N^{\text{vir}},\boldsymbol{Y}_N^{\text{vir}})$为第N组虚拟样本,初筛函数如下:

[0038]
$$\begin{aligned}&\phi_{\text{MMD}}((\boldsymbol{X}_1^{\text{vir}},\boldsymbol{Y}_1^{\text{vir}}),(\boldsymbol{X}_2^{\text{vir}},\boldsymbol{Y}_2^{\text{vir}}),\cdots,(\boldsymbol{X}_N^{\text{vir}},\boldsymbol{Y}_N^{\text{vir}}))=\\&\min(\text{MMD}(\boldsymbol{X}_1^{\text{vir}},\boldsymbol{X}_{\text{real}}),\text{MMD}(\boldsymbol{X}_2^{\text{vir}},\boldsymbol{X}_{\text{real}}),\cdots,\text{MMD}(\boldsymbol{X}_N^{\text{vir}},\boldsymbol{X}_{\text{real}}))\end{aligned}\tag{5}$$

[0039] 其中,min(•)表示取N组虚拟样本与$(X_{\text{real}},Y_{\text{real}})$MMD值最小的那组虚拟样本,即其作为质量最好的初筛虚拟样本$(\boldsymbol{X}_{\text{vir}}^{\text{inisel}},\boldsymbol{Y}_{\text{vir}}^{\text{inisel}})$。

[0040] 4.2.2基于PCA的虚拟样本可视化模块

[0041] 本申请中的DXN样本为高维样本,难以对所生成的虚拟样本的分布进行直观地感受。因此,本申请采用PCA将虚拟样本降到1维进行可视化以提供整体分布信息

[0042] PCA通过一组正交向量将原始数据投影到新的空间,消除了原始数据冗余的同时保留了主要信息。基于PCA的虚拟样本可视化实现步骤如下。

[0043] 首先,对$\boldsymbol{X}_{\text{vir}}^{\text{inisel}}$进行中心化处理,得到中心化样本U,其中$N_{\text{vir}}^{\text{inisel}}$为样本数量,$M_{\text{vir}}^{\text{inisel}}$为样本维数;

[0044] 接着,计算U的协方差矩阵C:

[0045] $C=UU^{T}$ (6)

[0046] 然后,采用特征分解法计算C的特征向量和特征值:

[0047] $C=W\Lambda W^{T}$ (7)

[0048] 其中,W为特征向量组成的矩阵;Λ为特征根按照从大到小顺序排列的对角阵;

[0049] 再然后,将$\boldsymbol{X}_{vir}^{inisel}$降到1维,如下:

[0050] $X_{PCA}=\mu_1 U$ (8)

[0051] 式中,X_{PCA}为降到1维的虚拟样本;μ_1为最大特征值对应的特征向量。

[0052] 最后,计算X_{PCA}分布函数,得到PCA可视化结果。

[0053] 4.2.3可视化分布信息判别模块

[0054] 虚拟样本PCA可视化结果提供的整体分布信息D_{PCA}:

[0055] $D_{PCA}=(R_{real}\cap R_{vir})/S_{real}$ (9)

[0056] 其中,R_{real}为真实样本分布函数与x轴包含的区域;R_{vir}为虚拟样本分布函数与x轴包含的区域;S_{real}为真实样本分布函数与x轴包含的区域的面积;$R_{real}\cap R_{vir}$表示R_{real}与R_{vir}重叠部分的面积。

[0057] 本申请提出的可视化分布信息判别函数如下:

[0058] $$\phi_{visual}(D_{PCA})=\begin{cases}1\ ,\ if\ D_{PCA}\geq\theta_S\\0\ ,\ else\ D_{PCA}<\theta_S\end{cases}\quad(10)$$

[0059] 其中,θ_S为依据经验设定的阈值。

[0060] 若$\Phi_{visual}(D_{PCA})$的值为1,表示该虚拟样本为合格虚拟样本;反之,该虚拟样本为不合格虚拟样本。

[0061] 4.3基于混合样本的风险预警模型构建模块

[0062] 判别得到的合格虚拟样本$(\boldsymbol{X}_{vir}^{fine},\boldsymbol{Y}_{vir}^{fine})$和真实样本$(X_{real},Y_{real})$组合,得到混合样本$S_{mix}$。

[0063] $$\boldsymbol{S}_{mix}=\{(\boldsymbol{X}_{vir}^{fine},\boldsymbol{Y}_{vir}^{fine}),(\boldsymbol{X}_{real},\boldsymbol{Y}_{real})\}\quad(11)$$

[0064] 使用随机森林(Random Forest,RF)作为风险预警模型的分类器,步骤如下。

[0065] 首先,利用Bootstrap算法和RSM算法对S_{mix}进行样本和特征的随机采样,获得N个子样本集;

[0066] 接着,利用N个子样本集构建N个决策树,每个决策树得到一个分类结果;

[0067] 最后,对N个分类结果进行投票,选择投票数量最多的类别作为最终的分类结果$\hat{\boldsymbol{Y}}$。

附图说明

[0068] 图1基于炉排炉的MSWI工艺流程图

[0069] 图2基于主动学习机制GAN的DXN排放风险预警模型构建策略

[0070] 图3基于GAN的虚拟样本生成流程图

[0071] 图4基于主动学习机制的虚拟样本评估和筛选流程

[0072] 图5基于DXN数据生成虚拟样本质量和epoch的关系

[0073] 图6DXN数据集50次风险预警模型的精度

[0074] 图7DXN数据测试风险预警实验结果

具体实施方式

[0075] 本申请所采用的DXN数据来自北京某基于炉排炉的MSWI电厂，涵盖了2012～2018年所记录的67个有效DXN排放浓度检测样本；原始输入特征经过处理后从314维降至120维，此处将输出DXN排放浓度分为5个风险等级，其划分标准如表1所示，其中，高风险、中高风险、中风险、中低风险和低风险相应的样本数为27、12、11、11和6。随机选择2/3作为训练集构建模型，剩下的1/3用于测试模型性能。

[0076] 表1 DXN排放风险等级划分标准

[0077]

分级标准	风险等级
$0.08 \le c(DXN)$	高风险
$0.04 \le c(DXN) < 0.08$	中高风险
$0.04 \le c(DXN) < 0.06$	中风险
$0.02 \le c(DXN) < 0.04$	中低风险
$c(DXN) < 0.02$	低风险

[0078] 对于DXN数据集：生成器隐含层采用Relu激活函数，输出层采用线性激活函数；判别器隐含层采用Relu激活函数，输出层采用Sigmoid激活函数，具体参数设置如表2所示。

[0079] 表2 DXN数据集虚拟样本生成实验参数设置

[0080]

数据集	生成器	判别器	批大小	学习率	训练代数
DXN	[121,600,120]	[121,600,1]	9	0.0001	2000

[0081] 图5表示基于DXN数据生成虚拟样本质量和epoch的关系。

[0082] 由图5可知，当epoch达到1000时，生成虚拟样本的质量达到稳定。因此，从1100次到2000次训练中每100次选择一个生成器，共10个生成器，每个生成器生成10组虚拟样本集，每组虚拟样本的5个风险等级各60个，共300个虚拟样本。从10个生成器的10组虚拟样本集筛选出与真实样本MMD值最低的作为初筛后的虚拟样本集。实验结果如表3所示。

[0083] 表3 DXN数据集基于MMD的虚拟样本初筛实验结果

[0084]

epoch	次数									
	1	2	3	4	5	6	7	8	9	10
1100	0.4470	0.4552	0.4652	0.4522	0.4580	0.4567	0.4706	0.4468	0.4685	0.4562
1200	0.4490	0.4580	0.4467	0.4481	0.4466	0.4620	0.4474	0.4556	0.4559	0.4660
1300	0.4429	0.4545	0.4474	0.4486	0.4494	0.4443	0.4544	0.4479	0.4540	0.4433
1400	0.4372	0.4372	0.4382	0.4298	0.4439	0.4345	0.4539	0.4340	0.4358	0.4348
1500	0.4354	0.4433	0.4347	0.4427	0.4428	0.4428	0.4339	0.4378	0.4375	0.4356
1600	0.4391	0.4415	0.4357	0.4306	0.4397	0.4362	0.4369	0.4328	0.4402	0.4395
1700	0.4386	0.4387	0.4332	0.4383	0.4341	0.4304	0.4346	0.4333	0.4485	0.4367
1800	0.4355	0.4367	0.4401	0.4346	0.4357	0.4387	0.4347	0.4422	0.4390	0.4345
1900	0.4380	0.4366	0.4340	0.4290	0.4330	0.4338	0.4374	0.4365	0.4347	0.4355
2000	0.4292	0.4276	0.4309	**0.4256**	0.4294	0.4272	0.4340	0.4299	0.4261	0.4332

[0085] 由表3可知，第2000次训练得到的生成器生成的第4组虚拟样本与真实样本的MMD值最低，因此选择该组虚拟样本。

[0086] 为保证可视化的效果，从300个虚拟样本随机选择27个低风险、12个中低风险、11个中风险、11个中高风险、6个高风险，共67个虚拟样本进行可视化。依据经验将阈值设为0.8，可视化结果得到的分布信息为0.81大于设定阈值。因此，该组虚拟样本为合格虚拟样本。

[0087] 使用上述合格虚拟样本和真实样本组成混合样本构建风险预警模型，相关参数如表4所示。

[0088] 表4 DXN数据集混合样本风险预警模型构建的相关参数

[0089]

真实样本数量	虚拟样本数量	RF 树的数量	样本划分
67	0	20	真实样本 2/3 用于训练，1/3 用于测试
67	67	20	混合样本 2/3 用于训练，1/3 用于测试

[0090] 50次实验的精度如图6所示。

[0091] 由图6可知，混合样本训练的风险预警模型性能好于真实样本训练的模型。

[0092] 此外，共进行5组对比实验，相关参数如表5所示。

[0093] 表5 DXN数据集基于混合样本的风险预警模型构建相关参数

[0094]

实验编号	样本数	各风险等级样本数	样本划分
A	67	[6 11 11 12 27]	真实样本 2/3 用于训练，1/3 用于测试
B	67	[6 11 11 12 27]	不平衡虚拟样本 2/3 用于训练，使用实验 A 的测试集测试
C	70	[14 14 14 14 14]	平衡虚拟样本 2/3 用于训练，使用实验 A 的测试集测试
D	134	[12 22 22 24 54]	不平衡混合样本 2/3 用于训练，1/3 用于测试
E	140	[28 28 28 28 28]	衡混合样本 2/3 用于训练，1/3 用于测试

[0095] 表5中，风险等级按照高风险、中高风险、中风险、中低风险和低风险的顺序排列。虚拟样本从筛选虚拟样本中随机抽取，其中不平衡虚拟样本和不平衡混合样本指其中各风险等级样本比例与真实样本的比例相同，平衡虚拟样本和平衡混合样本指其各风险等级的样本数相同。

[0096] 考虑RF算法的随机性，5种实验均重复执行50次。图7为实验A、B、C、D和E所构建的风险预警模型的准确率，表6得出了统计结果的对比。

[0097] 表6 DXN数据测试风险预警统计结果对比

[0098]

实验编号	准确率	
	平均（%）	标准差
A	48.9091	5.7759
B	48.0909	6.3017
C	47.4989	5.1204
D	70.8444	4.0065
E	78.8085	2.1901

[0099] 由上可知：1)真实样本的平均准确率为48.9091%，不平衡虚拟样本的平均准确率为48.0909%，平衡虚拟样本的平均准确率为47.4989%，本申请所提方法生成的虚拟样本很接近真实样本；2)基于混合样本的平均准确率为70.8444%和78.8085%，相较于未添加虚拟样本的准确率提升了44%和59%，表明添加虚拟样本有助于提高模型性能；3)平衡混合样本的平均准确率相较于不平衡混合样本提高了11%，表明平衡数据建模效果好于不平衡数据；4)混合样本准确率的标准差低于真实样本，表明添加虚拟样本有助于提高模型的稳定性。

[0100] 本申请提出基于可视化分布GAN的MSWI过程DXN排放风险预警方法，创新性表现在：1)首次提出基于GAN和可视化分布的DXN排放浓度风险预警策略；2)基于GAN的VSG方法

可以通过条件信息生成指定类型的虚拟样本,有效的扩展样本数量,填补真实样本的信息空白;3)基于可视化分布信息的虚拟样本评估和筛选方法使用MMD对虚拟样本进行初筛,对初筛虚拟样本可视化结果提供的分布信息进行判别,判别通过后得到合格的虚拟样本,合格虚拟样本的质量更加接近真实样本。基于工业DXN数据验证了所提策略和方法的有效性。未来研究方向包括:如何处理高维、离散的过程数据,如何使生成器和判别器在博弈对抗的过程中更加稳定,以获得更优质的虚拟样本。

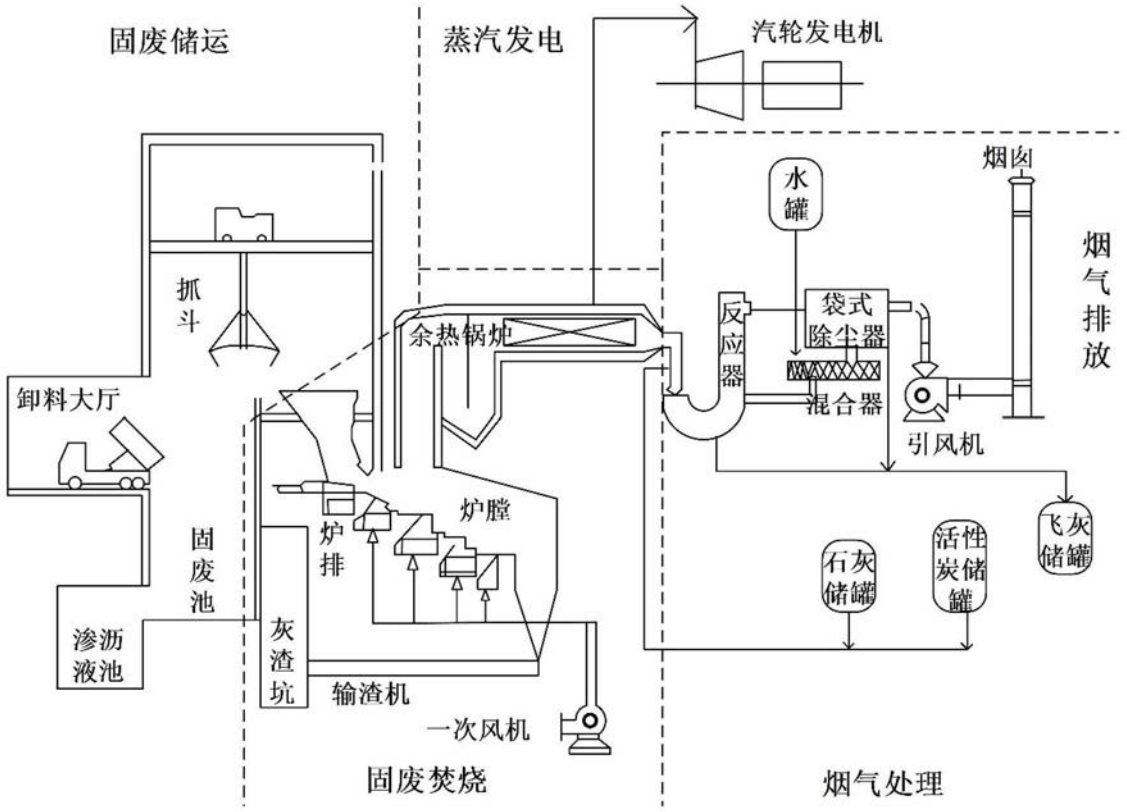

固废储运
蒸汽发电
汽轮发电机
烟囱
水罐
烟气排放
抓斗
余热锅炉
反应器
袋式除尘器
混合器
卸料大厅
引风机
炉膛
炉排
固废池
飞灰储罐
石灰储罐
活性炭储罐
渗沥液池
灰渣坑
输渣机
一次风机
固废焚烧
烟气处理

图1

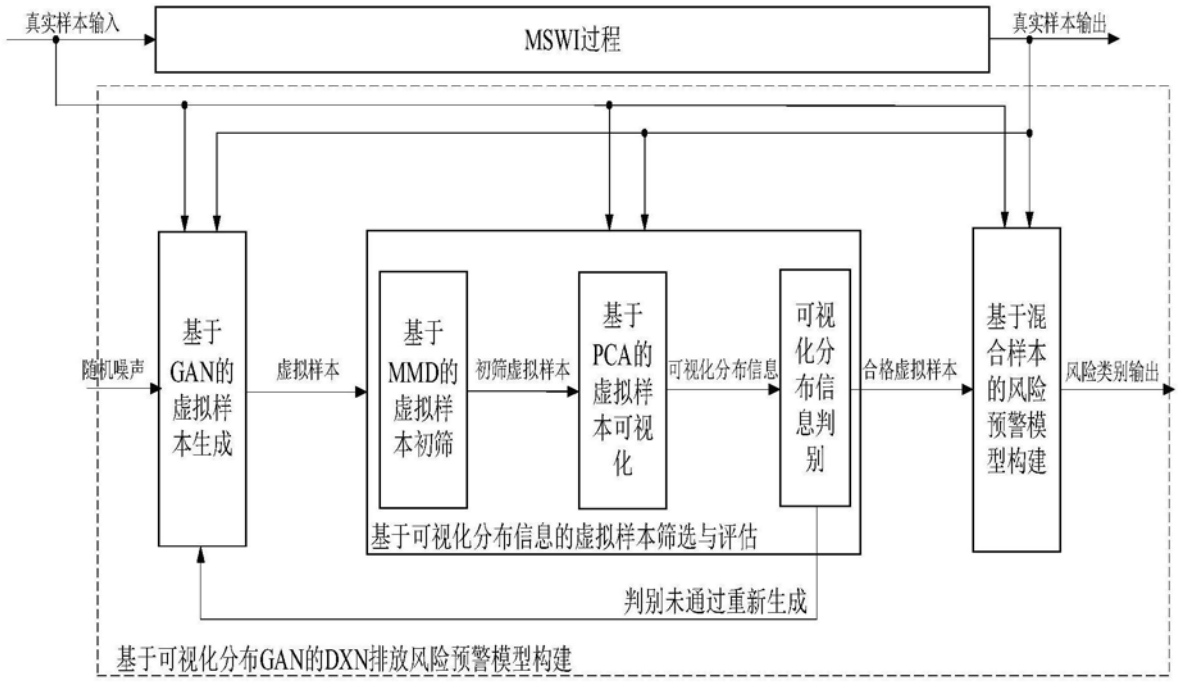

真实样本输入
MSWI过程
真实样本输出
随机噪声
基于GAN的虚拟样本生成
虚拟样本
基于MMD的虚拟样本初筛
初筛虚拟样本
基于PCA的虚拟样本可视化
可视化分布信息
可视化分布信息判别
合格虚拟样本
基于混合样本的风险预警模型构建
风险类别输出
基于可视化分布信息的虚拟样本筛选与评估
判别未通过重新生成
基于可视化分布GAN的DXN排放风险预警模型构建

图2

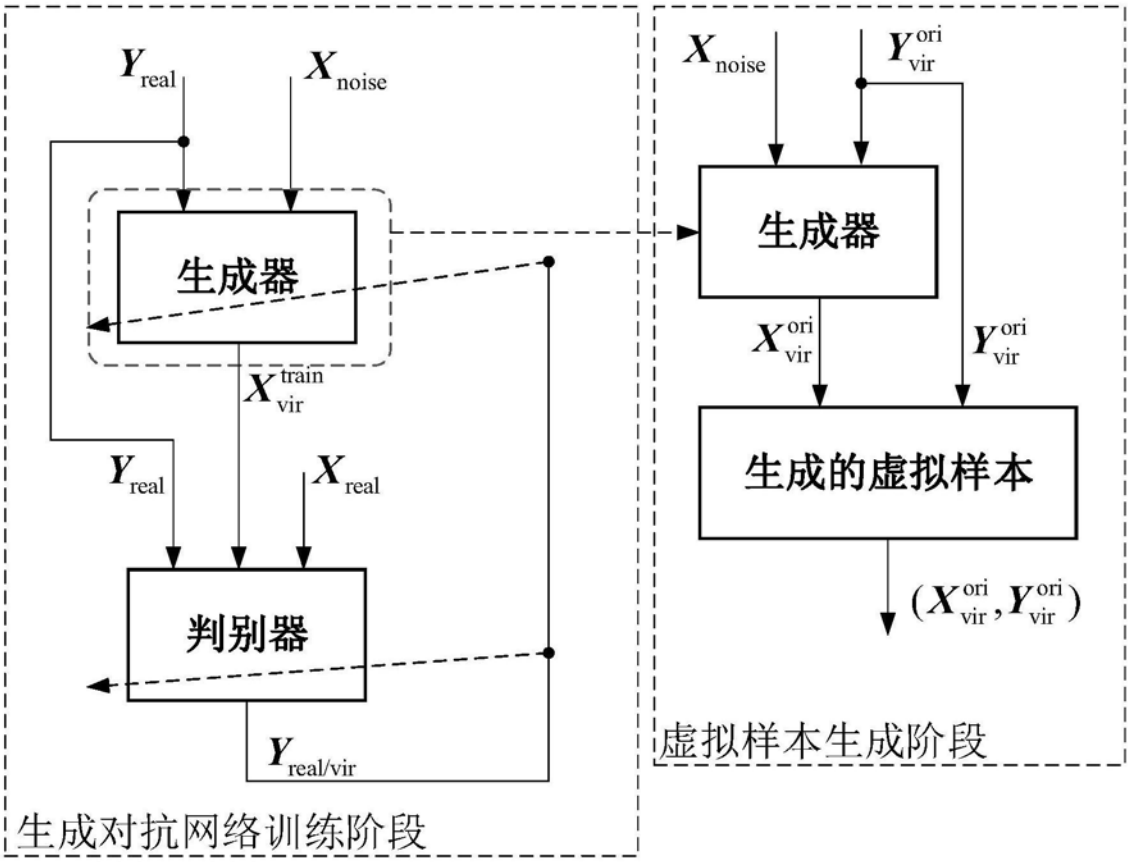

Y_{real}
X_{noise}
生成器
X_{vir}^{train}
Y_{real}
X_{real}
判别器
$Y_{real/vir}$
生成对抗网络训练阶段
X_{noise}
Y_{vir}^{ori}
生成器
X_{vir}^{ori}
Y_{vir}^{ori}
生成的虚拟样本
$(X_{vir}^{ori}, Y_{vir}^{ori})$
虚拟样本生成阶段

图3

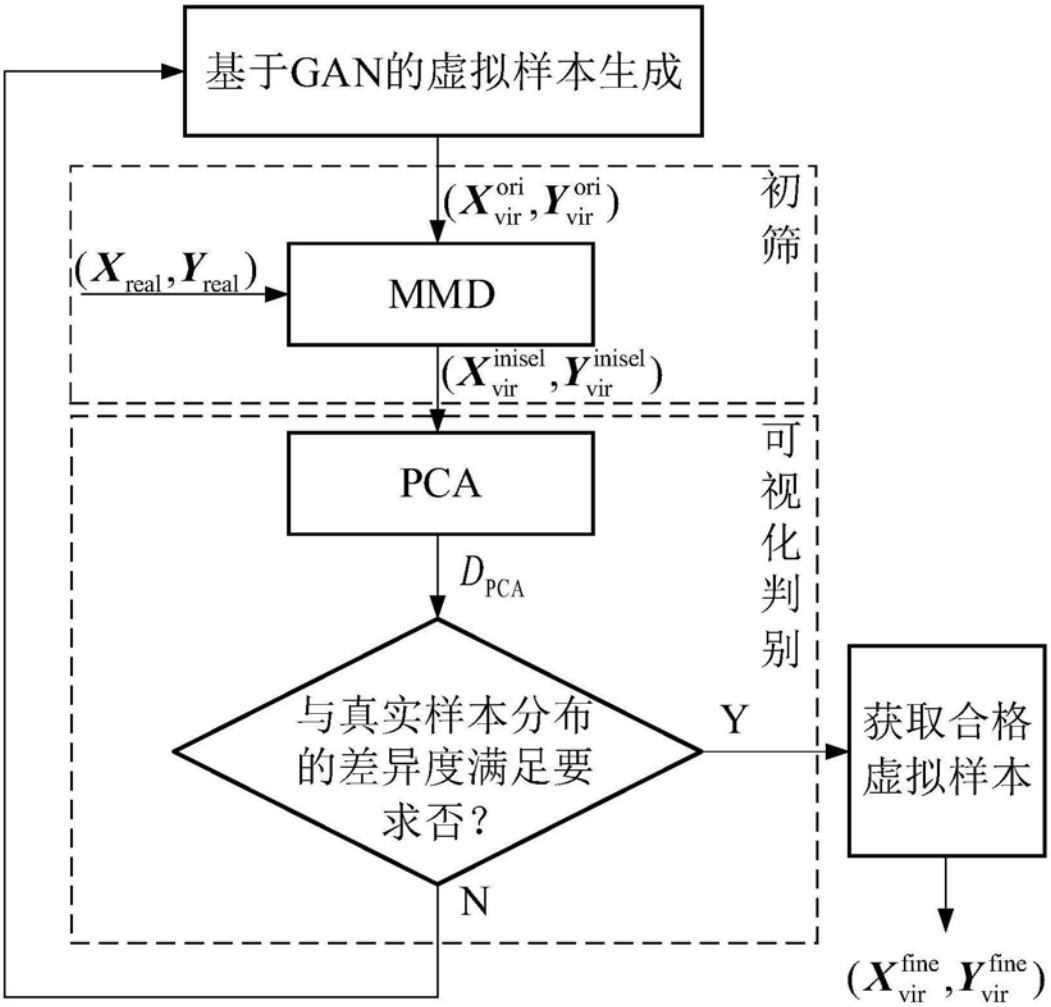

基于GAN的虚拟样本生成
$(X_{vir}^{ori}, Y_{vir}^{ori})$
初筛
(X_{real}, Y_{real})
MMD
$(X_{vir}^{inisel}, Y_{vir}^{inisel})$
PCA
可视化判别
D_{PCA}
与真实样本分布的差异度满足要求否？
Y
N
获取合格虚拟样本
$(X_{vir}^{fine}, Y_{vir}^{fine})$

图4

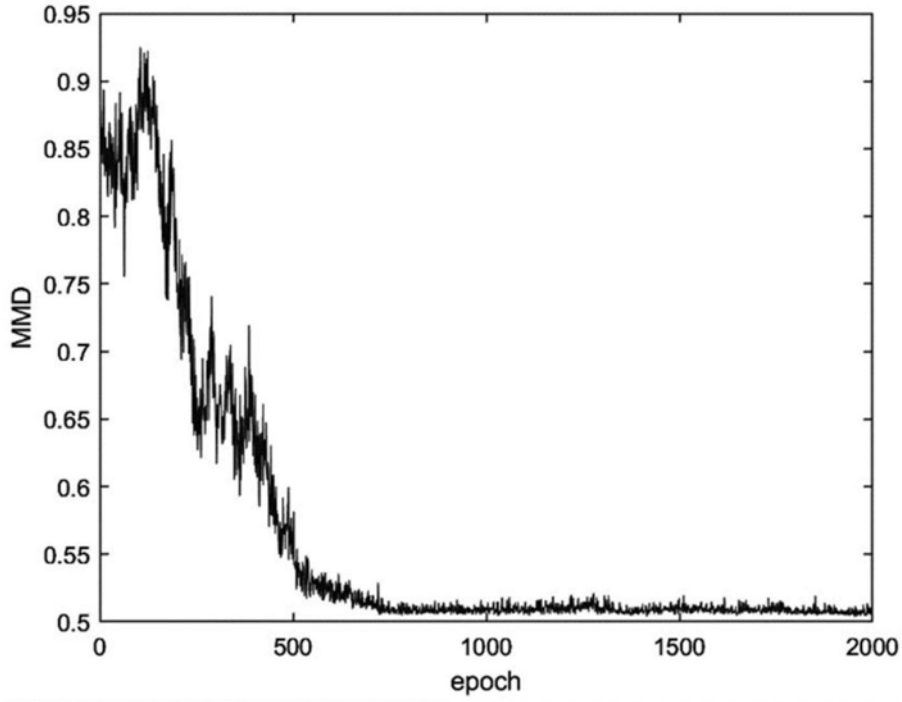

0.95
0.9
0.85
0.8
0.75
0.7
0.65
0.6
0.55
0.5
MMD
0
500
1000
1500
2000
epoch

图5

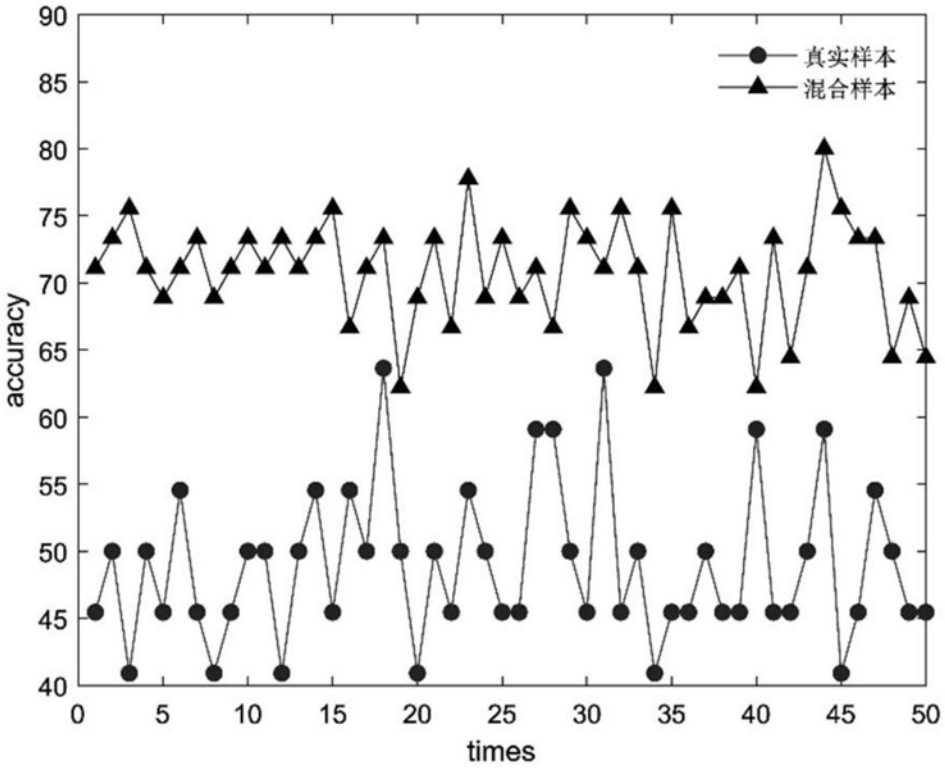

90
85
80
75
70
65
60
55
50
45
40
accuracy
0
5
10
15
20
25
30
35
40
45
50
times
真实样本
混合样本

图6

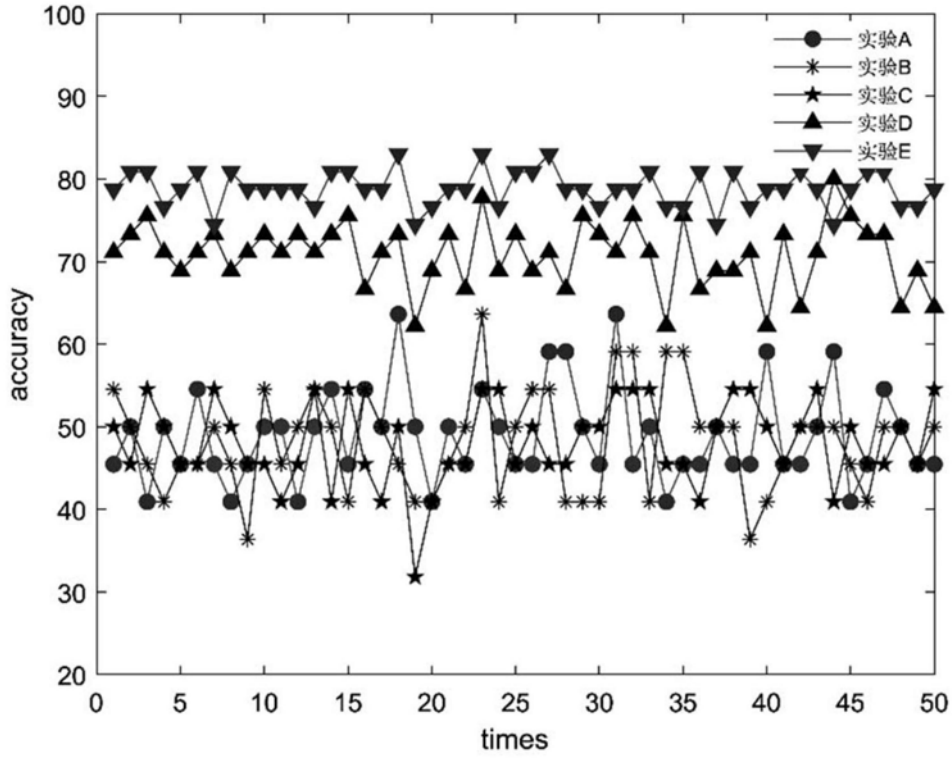

实验A
实验B
实验C
实验D
实验E
accuracy
times
100
90
80
70
60
50
40
30
20
0
5
10
15
20
25
30
35
40
45
50

图7