

# 面向复杂工业过程的虚拟样本生成综述

汤健<sup>1,2,3</sup> 崔璨麟<sup>1,2,3</sup> 夏恒<sup>1,2,3</sup> 乔俊飞<sup>1,2,3</sup>

**摘要** 用于复杂工业过程难测运行指标和异常故障建模的样本具有量少稀缺、分布不平衡以及内涵机理知识匮乏等特性。虚拟样本生成 (Virtual sample generation, VSG) 作为扩充建模样本数量及其涵盖空间的技术, 已成为解决上述问题的主要手段之一, 但已有研究还存在缺乏理论支撑、分类准则与应用边界模糊等问题。本文在描述复杂工业过程难测运行指标和异常故障建模所存在问题的基础上, 梳理虚拟样本定义及其内涵, 给出面向工业过程回归与分类问题的 VSG 实现流程; 接着, 从样本覆盖区域、实现流程与推广应用等方向进行综述; 然后, 分析讨论 VSG 的下一步研究方向; 最后, 对全文进行总结并给出未来挑战。

**关键词** 复杂工业过程, 虚拟样本生成, 数据驱动建模, 样本覆盖区域

**引用格式** 汤健, 崔璨麟, 夏恒, 乔俊飞. 面向复杂工业过程的虚拟样本生成综述. 自动化学报, 2024, 50(4): 688-718

**DOI** 10.16383/j.aas.c221006

## A Survey of Virtual Sample Generation for Complex Industrial Processes

TANG Jian<sup>1,2,3</sup> CUI Can-Lin<sup>1,2,3</sup> XIA Heng<sup>1,2,3</sup> QIAO Jun-Fei<sup>1,2,3</sup>

**Abstract** The modeling samples for difficulty to measure operation indexes and abnormal faults of complex industrial processes usually have the characteristics of sparse quantity, unbalanced distribution, and lack of connotation mechanism knowledge. Virtual sample generation (VSG) is a technology to expand the space and quantity of modeling samples and has become one of the main ways to solve the formerly mentioned difficulties. However, there are still some problems in the existing research results, such as the lack of theoretical support, the unclear of category criterion and the application boundary. First, the existing problems for difficulty to measure operational indexes and abnormal fault modeling of complex industrial processes are described. The definition of virtual samples and the connotation of virtual samples are combed, and the VSG implementation process for the regression and classification problems is provided. Second, the research status is summarized from the sample coverage area, implementation process, and application. Third, further research direction is analyzed and discussed. Finally, the summary and future challenges are given out.

**Key words** Complex industrial process, virtual sample generation (VSG), data-driven modeling, sample coverage area

**Citation** Tang Jian, Cui Can-Lin, Xia Heng, Qiao Jun-Fei. A survey of virtual sample generation for complex industrial processes. *Acta Automatica Sinica*, 2024, 50(4): 688-718

信息技术的不断发展和工业自动化进程的不断

深入, 利用多类型传感器采集的海量多模态数据能够支撑构建“工业大数据”驱动模型, 这已成为复杂工业过程实现智能控制、决策与优化的重要手段<sup>[1-4]</sup>。然而, 复杂工业过程的产品质量、污染物排放等难测关键运行指标和异常故障的建模数据依然存在量少稀疏、分布不平衡以及内涵机理知识匮乏等问题, 难以支撑构建准确且鲁棒的检测与识别模型<sup>[5-7]</sup>。以城市固废焚烧 (Municipal solid waste incineration, MSWI) 过程为例, 该过程排放的痕量有机污染物二噁英 (Dioxin, DXN) 因受限于在线检测技术的复杂度和离线化验技术的高成本, 使得具有真值的建模样本数量极少<sup>[8-9]</sup>; 此外, 已有的真值样本通常是在某种稳定的次优运行工况下获得的, 极优工况和潜在异常工况下的样本数据是缺失的。这些有限数量的真值样本中显然缺乏有助于洞悉运行指标的

收稿日期 2022-12-30 录用日期 2023-05-18

Manuscript received December 30, 2022; accepted May 18, 2023

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

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

本文责任编辑 谢永芳

Recommended by Associate Editor XIE Yong-Fang

1. 北京工业大学信息学部 北京 100124 2. 北京工业大学智慧环保北京实验室 北京 100124 3. 北京工业大学智能感知与自主控制教育部工程研究中心 北京 100124

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

相关机理,造成与建模相关的内涵知识匮乏。为解决上述问题,从扩增建模样本数量的视角,早期模式识别领域的研究学者 Poggio 和 Vetter 提出虚拟样本生成 (Virtual sample generation, VSG) 的概念<sup>[10]</sup>,其核心思想是基于已有数据通过某种方式生成并不存在的样本以扩充数据空间,其目前已广泛地应用于图像处理<sup>[11]</sup>、人脸识别<sup>[12]</sup>以及可靠性分析<sup>[13]</sup>等领域。图 1 给出了近 20 年内与 VSG 相关的文献发表数量与被引频次的变化情况。

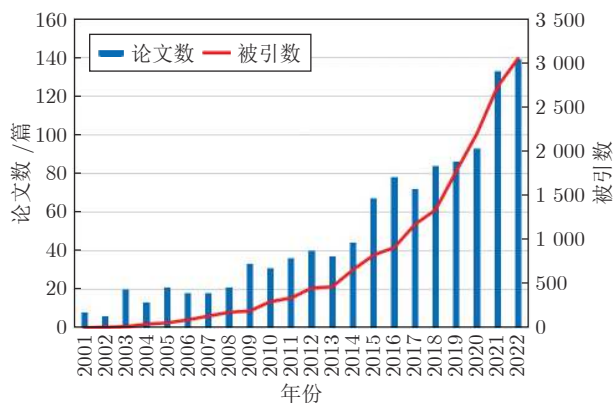


图 1 Web of Science 上的 VSG 论文数量与被引频次  
Fig. 1 Number and citation frequency of articles on VSG in Web of Science

由图 1 可知,有关 VSG 的论文发表量和被引量在总体上呈现上升趋势,表明该技术已逐渐受到研究学者的重视。虽然, Niyogi 等从数学视角证明了虚拟样本等价于将先验知识合并为正则化矩阵<sup>[14]</sup>,但复杂工业过程固有的机理不清、强耦合和非线性等特性,使得从该类过程获得明确的先验知识存在难度大和耗时长等问题,这导致目前研究学者大多聚焦于如何从小样本中学习知识进而生成虚拟样本的研究<sup>[15]</sup>。随着变分自编码器 (Variational autoencoder, VAE)<sup>[16-18]</sup>、生成对抗网络 (Generative adversarial network, GAN)<sup>[19-20]</sup>等生成模型的发展,使得 VSG 的研究热度得到进一步的提升<sup>[15]</sup>。随着工业数字孪生<sup>[21-22]</sup>、元宇宙<sup>[23-24]</sup>等概念的发展和日趋成熟,笔者认为, VSG 技术将成为上述技术发展中不可或缺的元素之一。综上, VSG 技术的逐步完善与成熟,能够为实际复杂工业过程的运行指标建模和异常故障识别乃至工业数字孪生和元宇宙提供有效支撑,有必要对当前 VSG 的研究动态与未来趋势进行总结与展望。

本文以工业过程为背景,全面综述 VSG 在工业过程中的研究现状及未来的发展方向,主要工作如下:第 1 节从样本稀缺、样本分布完备性差和样本内涵机理知识匮乏共 3 个视角总结工业过程

VSG 所面临的问题,并梳理虚拟样本定义、输入/输出空间虚拟样本内涵以及面向工业过程的实现流程;第 2 节根据目前的研究成果和实际工业过程的特点,从样本覆盖区域、实现流程与推广应用共 3 个方面进行综述;第 3 节给出相关的数据集和开源软件;第 4 节进行对比与讨论,并分析下一步的发展方向;第 5 节对全文工作进行总结并给出未来挑战。

## 1 面向工业过程的 VSG 技术

### 1.1 运行指标和异常故障建模存在的问题

目前,对系统性能、生产质量和经济效益的高要求使得现代工业过程的复杂度、包含的设备类型和数量也迅速增加,多类型传感器和自动化系统的应用促成了“工业大数据”以及工业过程建模、控制与优化研究<sup>[25]</sup>。相应地,基于数据驱动的运行指标和异常故障建模技术也得到迅速发展<sup>[26]</sup>。但是,技术上仍难以在线检测的部分运行指标和难以再现的异常故障却导致可用建模样本量稀缺的现象<sup>[27]</sup>。此外,复杂工业过程机理不清难以建模的特性和工业现场以确保安全稳定运行为目标的次优运行状态,使得建模数据还存在着分布不平衡以及内涵机理知识匮乏等问题。

#### 1.1.1 样本稀缺

针对难测运行指标而言,以 MSWI 过程的 DXN 排放浓度检测为例,其可采用离线直接检测法和在线间接检测法进行测量,但存在过程繁琐、价格昂贵、设备复杂和时间滞后等局限性;企业以月或季为间隔的不定期检测导致建模样本极为稀缺<sup>[8]</sup>。这需要采用适合于小样本数据的学习算法<sup>[28]</sup>。

针对复杂工业过程的故障检测与诊断 (Fault detection and diagnosis, FDD) 模型而言,异常故障样本属于“可遇不可求”,同时工业现场也是极力避免出现这样的故障,即会在异常出现前通过定期维修、降低生产效率等方式予以预防,因而导致样本缺失,增加了故障分类模型的构建难度<sup>[29-30]</sup>。

文献 [31] 指出,当在工程应用和学术研究中采用的建模样本数量分别少于 50 和 30 时,所面对的机器学习问题即被称为小样本学习问题;进一步,文献 [32] 将该问题表示为下式:

$$\alpha = \frac{n_{\text{sample}}}{p_{\text{feature}}} \quad (1)$$

式中,  $n_{\text{sample}}$  为样本数,  $p_{\text{feature}}$  为特征数,  $\alpha$  的典型取值为 {2, 5, 10}。显然,  $\alpha$  过小的数据集难以为构建可靠的学习模型提供支撑。

#### 1.1.2 样本分布完备性差

为保证工业全流程的运行安全性,实际工业过

程常工作在折衷的稳定状态, 甚至以牺牲经济性确保安全性为代价使工业过程长期运行在次优状态<sup>[33]</sup>. 因此, 即使采集了大量的过程数据, 但其所涵盖的工况波动范围和所具有的代表性样本数量也是有限的, 即多数为常规次优工况数据和少数为极优与潜在异常工况数据. 这些数据难以表征期望建模样本空间中所需要的完备分布. 本文将上述问题归纳为样本分布完备性差, 这会导致所构建的模型仅适用于稳定的次优运行过程, 难以适用于存在工况动态漂移变化的实际过程<sup>[34]</sup>.

在故障诊断中, 正常样本和异常故障样本间呈现的是长尾分布, 即正常运行与常见故障为头部多数类而罕见故障为尾部少数类, 这也是样本分布完备性差的体现, 其会严重影响故障诊断的结果. 度量少数异常类和多数正常类之间不平衡度的指标, 即不平衡比 (Imbalance ratio, IR)<sup>[35]</sup> 如下所示:

$$IR = \frac{N_{\text{majority}}}{N_{\text{minority}}} \quad (2)$$

式中,  $N_{\text{majority}}$  和  $N_{\text{minority}}$  分别为多数正常类和少数异常类样本的数量. 显然, IR 值越大表示建模样本集的不平衡程度越严重. 在文献 [36] 所构建的感应电动机故障诊断模型中, IR 的值达到了 10.

虽然目前已有针对少样本或零样本的故障诊断研究成果<sup>[37-39]</sup>, 但其在本质上并未解决样本分布完备性的问题.

### 1.1.3 样本蕴涵机理知识匮乏

用于难测运行指标与异常故障建模的过程数据所蕴涵的机理知识匮乏的原因在于: 首先, 样本数量稀缺; 其次, 样本分布不完备使得从数据中获取机理知识难, 尤其是在数据均源于单一工况的情况下; 再次, 工业过程的机理复杂不清导致知识理解难.

文献 [40] 指出, 针对在生产阶段早期采集的过程数据而言, 其所蕴涵的知识有限, 难以为推理样本分布提供支撑. 文献 [34] 认为, 虽然现代工业的规模在不断扩大, 但可用的过程信息却极为稀缺. 进一步, 文献 [41] 利用迁移学习从类似工况或设备的历史数据中获取知识, 将其用于当前过程关键工艺参数的预测; 文献 [42] 指出, 进行跨阶段 (Cross-phase)、跨状态 (Cross-state)、跨实体 (Cross-entity) 和跨领域 (Cross-domain) 的迁移学习, 是工业过程中获取知识的途径之一. 但是, 如何基于有限的建模样本和复杂工业过程的经验知识, 获得建模样本所蕴含的知识依然是一个开放性的问题.

## 1.2 虚拟样本的定义及内涵

### 1.2.1 虚拟样本的定义

虚拟样本的概念由 Poggio 和 Vetter 于 1992

年提出并用于模式识别领域<sup>[10]</sup>, 但并未给出明确定义. 文献 [43] 给出了如下所示的较为通用定义.

**定义 1.** 对于给定训练样本  $(\mathbf{x}_i, y_i)$ , 若由变换  $(T, f_T)$  得到的样本  $(\mathbf{x}'_i, y'_i)$  也是一个合理的样本, 那么就称新样本  $(\mathbf{x}'_i, y'_i)$  是由变换  $(T, f_T)$  所生成的虚拟样本.

基于定义 1, 文献 [6] 给出如下的推论:

**推论 1.** 给定由  $n$  个样本组成的训练样本集  $D = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$ , 通过某种合适的变换  $(T, f_T)$  可生成虚拟样本集合  $D' = \{(\mathbf{x}'_i, y'_i)\}_{i=1}^{n_{\text{vs}}}$ , 该过程表示如下:

$$\begin{cases} \{\mathbf{x}_i\}_{i=1}^n \xrightarrow{T} \{\mathbf{x}'_i\}_{i=1}^{n_{\text{vs}}}, \mathbf{x}'_{\text{low}} \leq \mathbf{x}'_i \leq \mathbf{x}'_{\text{high}} \\ \left. \begin{matrix} \{\mathbf{x}'_i\}_{i=1}^{n_{\text{vs}}} \\ \{y_i\}_{i=1}^n \end{matrix} \right\} \xrightarrow{f_T} \{y'_i\}_{i=1}^{n_{\text{vs}}}, y'_{\text{low}} \leq y'_i \leq y'_{\text{high}} \end{cases} \quad (3)$$

式中,  $\mathbf{x}'_{\text{low}}$  与  $\mathbf{x}'_{\text{high}}$  和  $y'_{\text{low}}$  与  $y'_{\text{high}}$  分别为虚拟样本输入和输出空间的上界与下界.

### 1.2.2 虚拟样本输入空间内涵

由于难测运行指标和异常故障建模样本的分布完备性差, 即样本分布在某个或某几个区域, 导致样本间存在大量间隙, 因此需要考虑对原始域样本空间进行有效填充. 此外, 原始域样本空间之外也可能存在符合实际数据分布的扩展域, 需对原始域进行有效扩展, 但扩展后可能会超出完备域 (期望域) 样本空间. 从可视化的角度, 图 2 给出了二维平面内原始域、扩展域和完备域 (期望域) 样本空间之内的虚拟样本和真实样本的相互关系<sup>[44]</sup>.

由图 2 可知, 生成的虚拟样本共有 4 类: 1) 在原始域样本空间内部填补真实样本间间隙的合格虚拟样本; 2) 在原始域样本空间外完备域 (期望域) 样本空间内的扩展域空间的合格虚拟样本; 3) 在扩展域样本空间外、完备域 (期望域) 样本空间内的合格虚拟样本; 4) 在完备域 (期望域) 样本空间外需剔除的不合格虚拟样本.

进一步, 文献 [45] 给出了三维空间视角下的不同虚拟样本输入生成方法的局限性, 如图 3 所示.

在图 3 中, 标记的数字是真实和虚拟样本编号, 以数字“12”为例, 其表示虚拟样本 12 是在真实样本 1 和 2 的连线上生成的. 具体而言, 图 3(a) 所示为依据样本顺序采用线性连续插值法依次在真实样本间插值生成虚拟样本输入, 即其仅分布在真实样本输入的顺序连线上; 图 3(b) 所示为合成少数类过采样技术 (Synthetic minority over-sampling technique, SMOTE), 其表示随机选择两个真实样本并在其间进行线性插值的方式, 显然其丢失了真实样

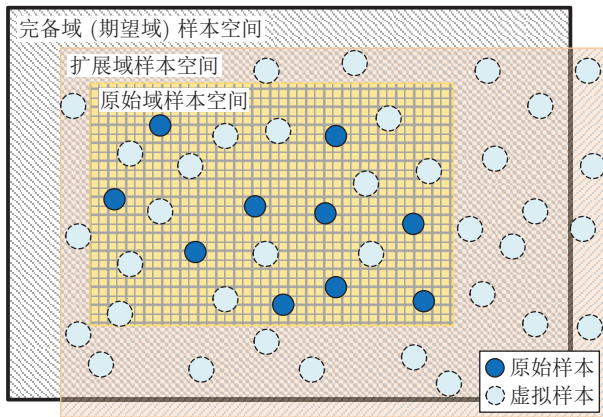


图 2 样本输入空间内虚拟与真实样本间的关系  
Fig.2 Relationship between virtual samples and real samples in sample input space

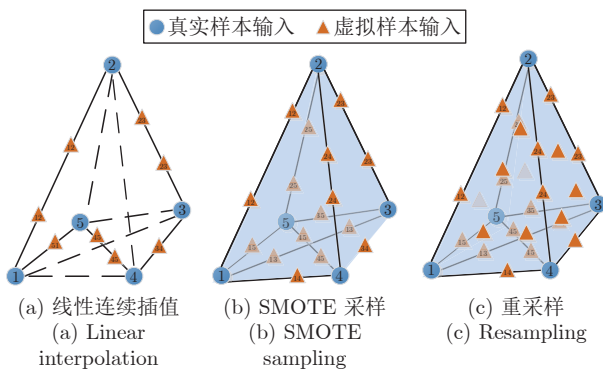


图 3 三维空间下的不同虚拟样本输入生成方法示意图  
Fig.3 Diagram of different virtual sample input generation methods in 3D space

本间可能存在的时序关系与物理含义；图 3(c) 所示为重采样法，其能够在真实样本的连接“面”上生成虚拟样本输入，但在由真实样本组成的空间内部并未生成虚拟样本，即存在样本“空洞”。

由上可知，虚拟样本输入的生成方式需要依据待解决问题而异，因此结合机理知识和经验知识是必要的。

### 1.2.3 虚拟样本输出空间内涵

针对样本输出空间而言，回归和分类问题具有完全不同的方式，下文分别描述。

#### 1) 回归问题

如何为虚拟样本输入匹配高精度的输出是面向回归的 VSG 需要面对的关键问题，其在极大程度上决定了虚拟样本的优劣。

目前，一般通过构建基于小样本的映射模型生成虚拟样本输出。Li 等<sup>[46]</sup>提出当映射模型的平均绝对百分比误差 (Mean absolute percentage error, MAPE) 不超过 10% 时，其可用于生成虚拟样本输出。

基于映射模型生成虚拟样本输出的流程如图 4 所示。

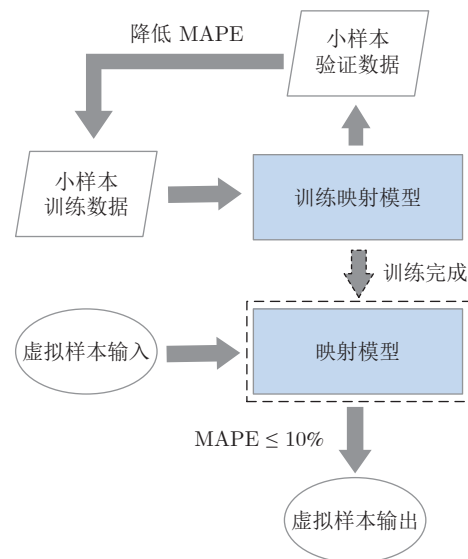


图 4 映射模型生成虚拟样本输出流程图  
Fig.4 Flow chart of virtual sample output generation based on mapping model

针对不同的映射模型结构，通过调整参数虽然可达到上述要求，但由于模型自身的差异性，由相同虚拟样本输入所映射的输出间也存在不同。因此，为得到最佳的虚拟样本输出，映射模型需对数据集具有较好的适应性。

#### 2) 分类问题

相较于回归问题，面向分类的虚拟样本输出所面临的问题是类间不平衡，即某些类的样本数量远少于另外一些类。

针对故障诊断模型而言，充足的训练样本和完备的故障类型是需要满足的两个基本条件<sup>[47]</sup>。受工业过程复杂性和检测环境不稳定性等限制，异常故障数据采集困难，某类故障甚至不可再现<sup>[48]</sup>。图 5 给出了多数正常类和少数异常类真实样本与虚拟样本间的关系。

由图 5 可知，面向分类问题的 VSG 的特点为：a) 数量少的类别 (少数类) 需要生成更多的虚拟样本，数量多的类别 (多数类) 只需生成少量甚至不生成虚拟样本；b) 少数类虚拟样本主要生成稀疏区域以填补信息空缺；c) 多数类和少数类都需要在分类边界上生成一定量的虚拟样本。此外，因工业过程的动态变化，还可能不存在不能采集到样本的未知类，这需要机理知识与经验知识支撑。

从本质上，回归问题和分类问题中的 VSG，都

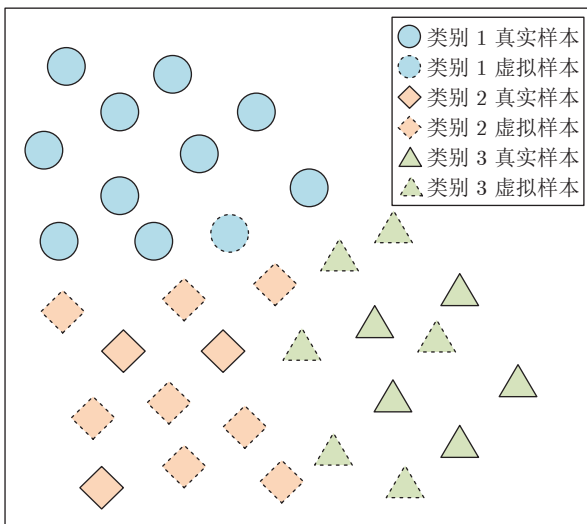


图 5 面向分类问题的虚拟与真实样本间的关系

Fig.5 Relationship between virtual samples and real samples for classification problem

很难从复杂工业过程获得清晰机理知识和领域先验知识. 从理论支撑方面而言, Niyogi 等通过数学推导证明了虚拟样本等价于合并先验知识以作为正则化矩阵<sup>[4]</sup>, 但是, 在期望分布、虚拟与真实样本相似度以及混合样本组成等方面的研究还缺乏理论支持.

### 1.3 面向工业过程的 VSG 实现流程

基于小样本建模数据的工业过程 VSG 实现流程如图 6 所示.

如图 6 所示, 步骤如下:

1) 第 1 阶段为过程数据预处理, 包括高维数据降维、缺失数据填补和过程数据标准化等操作以及机理与经验知识获取.

2) 第 2 阶段为生成虚拟样本输入, 对于回归问题而言要求能够填补完备域样本空间, 对于分类问题而言要求保证少数类和多数类间的平衡性.

3) 第 3 阶段为匹配虚拟样本输出, 对于分类问题而言, 因其类别标签是预设的和确定的而相对简单. 对于回归问题而言, 其输出真值需通过映射模型进行匹配而相对复杂. 但分类问题可能需要考虑未知类.

4) 第 4 阶段为生成虚拟样本质量筛选, 通过相似性度量以及建模结果误差等准则进行筛选以保证虚拟样本质量.

5) 第 5 阶段为生成虚拟样本数量确定, 通过获得理想的期望数量以减少计算成本和提高模型精度, 目前还缺少理论支撑.

在上述流程中, 第 1 阶段是 VSG 的必要操作, 第 2 和 3 阶段是 VSG 的基础操作, 第 4 和 5 阶段

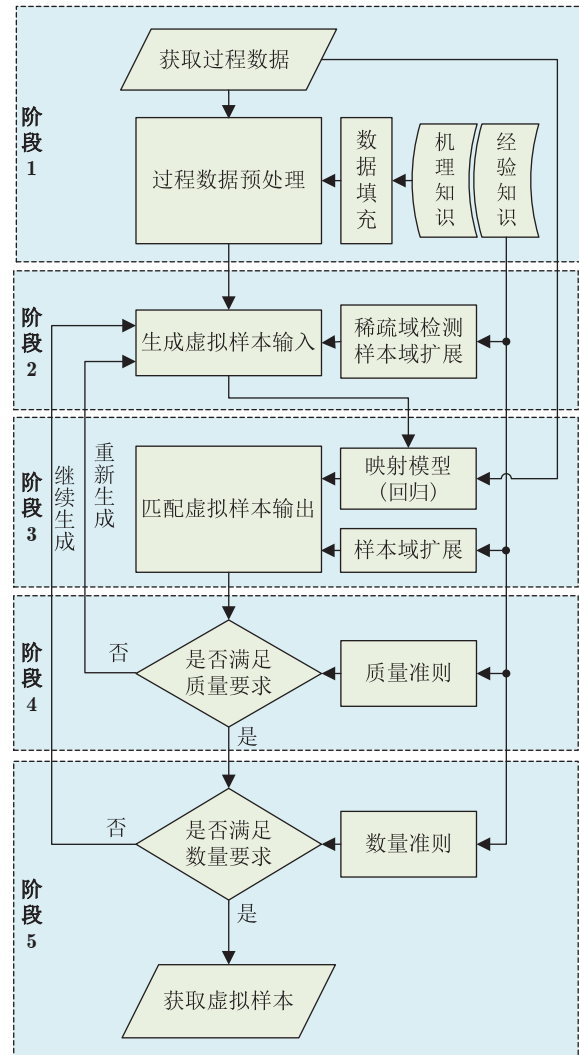


图 6 面向工业过程的 VSG 实现流程图

Fig.6 Flow chart of VSG for industrial process

是生成高质量虚拟样本的重要保障.

此外, 在已有研究成果中, 存在先进行阶段 3 再进行阶段 2 的 VSG 流程, 如文献 [20] 和文献 [49] 等. 这类方法相对较少, 本文在后文综述时也予以说明.

## 2 VSG 的研究现状

本节将面向工业过程数据的 VSG 研究现状从样本覆盖区域、实现流程和推广应用共 3 个方面进行综述, 之后针对每个方向再进行展开叙述, 具体如图 7 所示.

### 2.1 基于样本覆盖区域分类的研究现状

#### 2.1.1 基于原始域样本空间的 VSG

基于原始域样本空间的 VSG 通过挖掘原始样

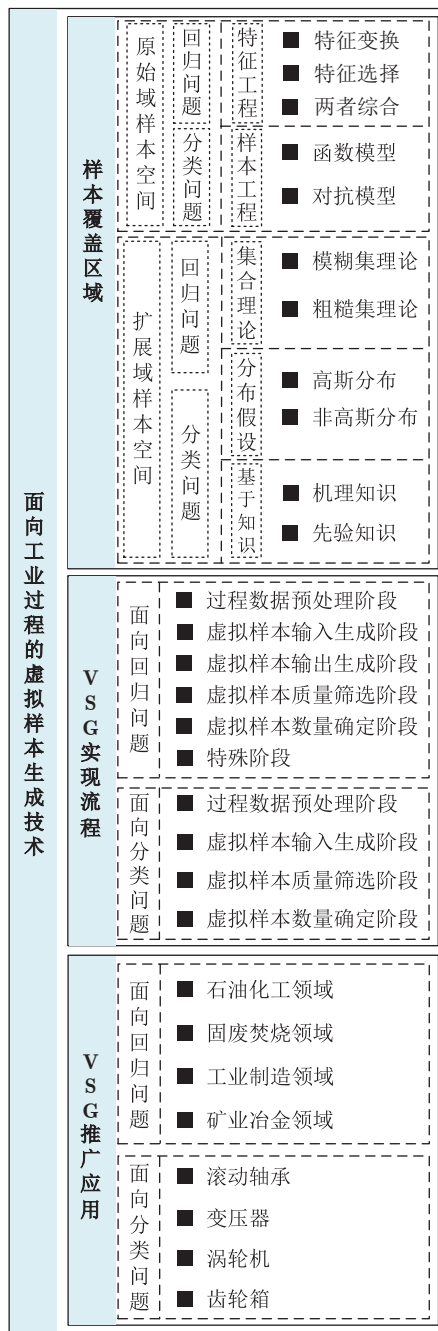


图7 VSG的研究现状结构图

Fig.7 Structure diagram of VSG research status

本间的分布关系以生成虚拟样本,其重点关注的是原始域样本空间的稀疏区域,目的是通过虚拟样本填补真实样本间的空隙。下文针对回归和分类问题分别从特征工程和样本工程2个视角进行描述。

### 1) 面向回归问题的VSG

#### a) 特征工程

复杂工业过程的运行指标建模数据具有高维度特性<sup>[50-51]</sup>,对应的稀疏区域难以识别,这导致直接进行VSG存在困难。因此,先进行特征工程是广泛采

用的解决方法。下面从特征变换、特征选择和两者综合共3个方面进行综述。

特征变换是指通过线性或非线性的方式将原始数据变换至新的低维或高维空间。Zhu等<sup>[52]</sup>先采用局部线性嵌入(Locally linear embedding, LLE)算法对高维数据进行降维,再基于随机插值生成虚拟样本输入,最后通过反向传播神经网络(Back propagation neural network, BPNN)映射模型得到虚拟样本输出。Zhang等<sup>[53]</sup>基于等间隔映射(Isometric feature mapping, Isomap)对高维数据进行可视化以寻找稀疏区域后采用插值法和映射模型生成虚拟样本。文献<sup>[54]</sup>采用 $t$ 分布随机邻域嵌入( $t$ -distributed stochastic neighbor embedding,  $t$ -SNE)算法,在提取原始高维特征后再插值生成虚拟样本输入,通过随机森林(Random forest, RF)映射模型得到虚拟样本输出。上述这些方法的本质是在变换后的特征空间中获得易生成虚拟样本的区域,但并未考虑原始特征中可能存在的冗余和变换后特征失去的原有物理含义等问题。

相较于特征变换,特征选择虽然会舍弃掉部分特征,但能够保留清晰的物理含义,更适合于在输入输出间具有较强因果关系的工业过程。陈忠圣等<sup>[55]</sup>基于精对苯二甲酸生产过程的机理,选择影响醋酸消耗的17个因素作为输入特征后采用分位数回归条件GAN生成虚拟样本。该方法适用于减少特征后可清晰地获得易生成区域的建模样本,但也存在约简后仍然难以分辨稀疏区域以及忽略的特征在未知工况下可能造成的未知影响等问题。

此外,现有研究成果中也存在串联特征变换和特征选择两种方式生成虚拟样本的策略<sup>[56]</sup>,该研究根据专家经验和MSWI过程DXN排放机理选择输入特征后再基于改进大趋势扩散和隐含层插值生成虚拟样本。这类方法需结合具体工业过程予以应用,具有较强的定制化特性。

#### b) 样本工程

样本工程旨在直接学习原始真实样本所表征的分布关系,基于样本“间隙”生成虚拟样本。根据所选用模型的不同,可分为基于函数模型插值和基于对抗模型生成两种方式。

基于函数模型插值的VSG是指通过某种函数表征原始真实样本的间隙,基于该函数生成虚拟样本输入后再结合映射模型生成虚拟样本输出。典型方法包括分段线性插值、径向基函数(Radial basis function, RBF)插值和三样条插值(Cubic spline interpolation, CSI)等。Zhu等<sup>[57]</sup>采用空间投影法进行稀疏性检测以得到原始样本间的空隙,利用中

点插值和 RBF 映射模型生成虚拟样本. 进一步, Chen 等<sup>[49]</sup> 基于稀疏性假设和中心假设确定虚拟样本数量, 基于 CSI 生成虚拟样本输出后再经过输入训练神经网络 (Input-training neural network, IT-NN) 获得虚拟样本输入, 结果表明可有效地提高模型性能. Sutojo 等<sup>[58]</sup> 采用总线拓扑结构, 在原始样本间连接后再在连接线上插值生成虚拟样本的策略.

目前, 在如何选取合适的用于产生虚拟样本输出的映射模型方面还不存在统一论. 相关研究包括: 通过随机权重神经网络 (Random weight neural network, RWNN) 模型学习样本间的非线性关系后在其隐含层插值以生成虚拟样本的策略<sup>[59]</sup>, 其首先在真实样本输出之间插值生成虚拟样本输出, 然后在隐含层插值得到新隐含层并反向求出虚拟样本输入, 最后组合虚拟样本输入和输出; 进一步, 朱宝等<sup>[60]</sup> 提出在自联想神经网络 (Auto-associative neural network, AANN) 的隐含层插值生成虚拟样本以消除样本间的噪声; 再随后, 乔俊飞等<sup>[66]</sup> 提出基于等间隔插值和正则化 RWNN 隐含层插值获取虚拟样本并删除冗余样本, 进而增强了虚拟样本的稳定性和互补性; 进一步, 汤健等<sup>[61]</sup> 提出基于粒子群优化 (Particle swarm optimization, PSO) 算法优化选择上述方法所生成的虚拟样本以降低虚拟样本之间的冗余性; 为了有效地均衡虚拟样本数量和模型泛化性能, 文献<sup>[15]</sup> 提出基于多目标 PSO 混合优化的 VSG, 其采用 RF 和 RWNN 集成模型作为非线性映射模型.

近些年, 深度学习在学术界发展迅速并在工业界广泛应用, 体现出极强的处理复杂任务的能力<sup>[62]</sup>. GAN 是目前深度学习中最为热门的研究方向之一<sup>[19]</sup>, 其虽已广泛应用于图像生成领域, 但在工业过程 VSG 中的研究才刚刚起步<sup>[63]</sup>. GAN 的基本原理是: 通过生成器和判别器的博弈对抗使得生成的虚拟样本越来越接近真实样本, 生成器的目标是生成判别器无法判别的样本, 判别器的目标是准确识别真实样本和虚拟样本, 其结构如图 8 所示<sup>[63]</sup>.

GAN 的目标函数表示如下:

$$\min_G \max_D V(D, G) = E_{x \sim p_{\text{data}}(x)} [\ln(D(x))] + E_{z \sim p_z(z)} [\ln(1 - D(G(z)))] \quad (4)$$

式中,  $p_{\text{data}}$  为原始小样本的分布;  $p_z$  为随机噪声的分布;  $D(x)$  和  $G(z)$  分别表示判别器和生成器的输出.

面向回归问题, 针对基于 GAN 的 VSG, 如何为其所生成的虚拟样本输入映射合理的输出是目前

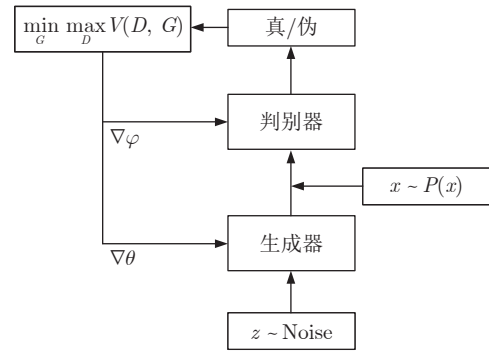


图 8 GAN 模型的结构

Fig.8 Structure of GAN model

的研究难题. 对此, Zhu 等<sup>[20]</sup> 通过计算局部异常因子 (Local outlier factor, LOF) 确定稀疏区域, 采用  $K$ -means++ 算法计算簇的中心后插值生成虚拟样本输出, 将其作为条件 GAN (Conditional GAN, CGAN) 的条件变量生成相应的虚拟样本输入; 在此基础上, 文献<sup>[64]</sup> 提出基于循环结构 CGAN (Cycle structure CGAN, CS-CGAN) 的 VSG, 采用最近邻距离确定离群点以获得稀疏区域边界, 通过 WGAN-GP 在稀疏区域生成虚拟样本输入, 之后利用 CS-CGAN 生成和选择虚拟样本输出; 进一步, He 等<sup>[65]</sup> 通过 GAN 内嵌分位数回归器生成与虚拟样本输入相匹配的虚拟样本输出. 上述方法均未考虑如何结合具体工业过程领域知识进行区域扩展和确定虚拟样本数量.

## 2) 面向分类问题的 VSG

### a) 特征工程

目前, 面向分类问题的 VSG 多应用于故障诊断领域, 采用特征工程进行处理的故障样本大多为机械振动信号. 这类 VSG 的特点是: 先采用快速傅里叶变换 (Fast Fourier transform, FFT) 将时域信号转换至频域再在生成模型中对特征进行处理, 如: 添加卷积层提取特征<sup>[66-67]</sup>、采用编码器提取特征<sup>[68]</sup> 以及添加自注意力模型增强特征<sup>[69]</sup> 等.

### b) 样本工程

从函数模型插值和对抗模型生成两个方面进行介绍. 相较于回归问题而言, 因无需考虑生成虚拟样本输出, 已有的面向分类问题的 VSG 更关注类与类之间的关系以及类间数据的平衡.

SMOTE 通过在邻近少数类样本间的随机线性插值生成少数类的虚拟样本, 进而实现不平衡数据集的均衡化<sup>[70]</sup>, 如下所示.

$$\mathbf{x}_{\text{vir}, i}^j = \mathbf{x}_i + \text{rand}(0, 1) \cdot (\hat{\mathbf{x}}_i^j - \mathbf{x}_i) \quad (5)$$

式中,  $\mathbf{x}_i$  为第  $i$  个少数类样本,  $\hat{\mathbf{x}}_i^j$  为  $\mathbf{x}_i$  的第  $j$  个  $K$

近邻样本,  $x_{\text{vir}, i}^j$  为生成的虚拟样本,  $\text{rand}(0, 1)$  为服从  $(0, 1)$  范围均匀分布的随机数。

SMOTE 可归类为基于分段线性插值的 VSG 方法. 在此基础上, Mathew 等<sup>[71]</sup> 提出基于加权核的 SMOTE, 其通过在支持向量机 (Support vector machine, SVM) 的特征空间中进行插值生成虚拟样本的方式解决算法在高 IR 下的非线性可分离问题; 进一步, Maldonado 等<sup>[72]</sup> 提出面向高维数据集的改进 SMOTE, 通过特征排序法选择相关特征后采用 Minkowski 距离替换欧氏距离以生成高维虚拟样本; 谢桦等<sup>[73]</sup> 先通过 SMOTE 生成虚拟样本, 再采用决策树算法提取有关变压器状态的评估知识; 随后, 刘云鹏等<sup>[74]</sup> 针对变压器非正常状态的样本数量稀少的问题, 提出基于 SVM 和 SMOTE 的变压器故障诊断方法, 其核心理念是在支持向量近似的分类边界上采用最近邻决策机制生成虚拟样本输入; Soltanzadeh 等<sup>[75]</sup> 针对噪声样本偏移和边界样本重叠等问题, 提出可以识别类类边界和控制生成范围的 SMOTE.

针对多类数据混杂问题, 文献 [76] 提出采用组发现技术对原始样本进行预分类以生成指定类的虚拟样本, 其过程为: 先任选样本点  $P_1$ , 将与  $P_1$  同类和不同类样本点间的最小距离记为  $R_1$ ; 接着再以  $P_1$  为球心和  $R_1$  为半径构建超球体; 之后进行判断, 若在超球体内存在与  $P_1$  同类的样本  $P_2$ , 则  $P_1$  和  $P_2$  为同组; 最后, 以  $P_2$  为球心重复上述操作, 直至超球体不包含新的同类样本, 进而完成预分类. 随后, 文献 [77] 在组发现技术的基础上采用纯化过程剔除相近的非同类样本以保证分组的准确性, 之后再通过构造超球以生成虚拟样本输入, 实验表明该策略能够有效地提高接地网络的故障识别率.

由于面向分类问题的虚拟样本输出为类别标签, 故其可作为已掌握的条件信息控制生成模型以获得指定类型的虚拟样本. 例如, 文献 [78] 将少数类的标签作为条件信息输入 CGAN, 结构如图 9 所示.

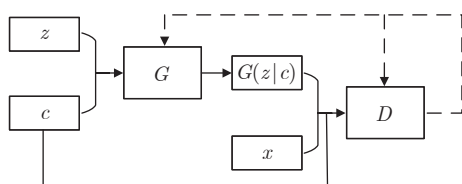


图 9 基于 CGAN 的 VSG 模型结构

Fig.9 VSG model structure based on CGAN

在图 9 中, 随机噪声  $z$  和类别标签  $c$  共同注入生成器  $G$ , 其中  $c$  作为条件信息控制  $G$  生成对应的虚拟样本输入  $G(z|c)$ ; 将真实样本输入  $x$  和虚拟样

本输入  $G(z|c)$  注入判别器  $D$  后根据判别结果更新  $D$  和  $G$ .

基于上述模型结构, 黄南天等<sup>[79]</sup> 构建基于辅助分类器 GAN (Auxiliary classifier GAN, ACGAN) 的风机主轴承故障诊断模型, 提出通过添加 Dropout 层防止过拟合以减少重复样本生成的策略; Li 等<sup>[80]</sup> 采用具有梯度惩罚的辅助分类 Wasserstein GAN (Auxiliary classifier Wasserstein GAN with gradient penalty, ACWGAN-GP) 生成具有高质量的少数类虚拟样本以提高模型准确率; Dixit 等<sup>[81]</sup> 提出采用模型无关元学习 (Model agnostic meta learning, MAML) 算法替换常规的随机梯度下降算法进而初始化和更新网络参数的条件辅助分类 GAN, 提高了生成模型的稳定性; Yang 等<sup>[82]</sup> 采用基于 GAN 的 VSG 解决谐波传动故障数据不平衡问题后利用多尺度卷积神经网络 (Convolutional neural network, CNN) 进行故障诊断; Wang 等<sup>[67]</sup> 采用深度卷积生成对抗网络 (Deep convolutional GAN, DCGAN) 生成虚拟样本平衡训练集后通过  $K$ -means 聚类算法构建改进 CNN 诊断模型; Zareapoor 等<sup>[83]</sup> 提出采用判别器既判断样本真假又充当分类器和故障检测器的少数类过采样 GAN (Minority oversampling GAN, MoGAN) 策略, 有效地提高了虚拟故障样本的质量; 之后, Li 等<sup>[84]</sup> 和 Li 等<sup>[85]</sup> 对 WGAN 进行改进以稳定生成的故障样本质量; 李东东等<sup>[86]</sup> 基于贝叶斯优化策略自适应调节 GAN 的判别器参数和采用 Wasserstein 距离作为损失函数提高模型的泛化性能, 结果表明其能够有效提高虚拟样本的质量; 此外, 也有研究人员组合多个 GAN 进行 VSG 后, 再通过筛选提高虚拟样本的质量<sup>[82]</sup>. 在上述研究中, 仅是考虑了依据已知的类别生成虚拟样本输入, 但如何面对动态环境下的未知类别进行 VSG 还有待于研究.

GAN 的本质是基于博弈对抗的训练框架, 其能够训练任意类型的生成模型. 自编码器 (Autoencoder, AE) 作为一种非线性无监督神经网络<sup>[87]</sup>, 通过非线性变换将输入数据投影至潜在特征空间中. 变分 AE (Variational AE, VAE) 是以 AE 结构为基础的深度生成模型<sup>[88]</sup>. 将编码器与 GAN 进行组合可得到基于编码器的 GAN, 其在 VSG 领域的研究成果包括: 戴俊等<sup>[89]</sup> 将 AE 的解码器嵌入至 GAN 中作为生成器, 并通过编码-解码-编码过程后的特征差异判断是否存在异常; Wang 等<sup>[68]</sup> 建立基于条件变分自编码器 GAN (Conditional VAE GAN, CVAE-GAN) 的不平衡故障诊断模型, 通过 CVAE 获取故障样本分布作为生成器输入, 利用博弈对抗机制对生成器、判别器和分类器的参数进行优化;

Liu 等<sup>[90]</sup> 将编码器合并到 GAN 中, 通过学习真实数据的深度特征以提高数据生成质量, 通过深度遗憾分析算法对判别器施加梯度惩罚以避免模式崩溃, 实验表明具有较好的鲁棒性; Wang 等<sup>[91]</sup> 设计具有传输层的改进型 AE 以消除数据噪声, 采用暹罗编码器结构计算潜在特征之间的残差, 引入最小二乘 GAN (Least squares GAN, LSGAN) 学习健康数据分布以生成虚拟样本, 结果表明可提前检测潜在异常; Liu 等<sup>[92]</sup> 将自调制嵌入到 GAN 的生成器中, 使其能够同时依靠输入和判别器反馈进行参数更新; Rathore 等<sup>[93]</sup> 提出结合堆叠 AE 和 WGAN 的 VSG 策略, 提高了虚拟样本的质量. 可见, 如何获得具有可解释性的 GAN 还有待深入研究, 例如基于模糊或决策树算法. 此外, 面向回归问题的基于编码器的 GAN 还有待进一步研究.

### 2.1.2 基于扩展域样本空间的 VSG

理论上, 由真实小样本组成的原始域样本空间是完备 (期望) 域样本空间的子集<sup>[44]</sup>, 其蕴含信息有

限. 从实际工业过程视角, 所采集的真实小样本多源自于某种平稳工况, 但完备域样本空间需要同时覆盖平稳与非平稳工况. 因此, 研究人员开始关注在原始域样本空间 (易获取数据) 上进行扩展以得到扩展域样本空间 (难获取数据), 并在其上生成虚拟样本, 进而能够接近完备域样本空间<sup>[15]</sup>. 理论上, 扩展域可分为可扩展域和未知域, 后者无数据可用, 即不存在真值或是未知类别. 针对工业过程的多输入单输出回归和分类问题, 面向 VSG 的原始域、可扩展域和未知域的示意图如图 10 所示.

笔者将基于扩展域样本空间的 VSG 分为面向集合理论、面向分布假设和基于知识共 3 类, 从回归和分类 2 个方面进行综述.

#### 1) 面向回归问题的 VSG

##### a) 集合理论

工业过程的真实小样本携带的有限信息导致进行 VSG 存在不可避免的不确定性. 模糊集理论是处理具有随机和不确定特性数据的有效手段. 鉴于

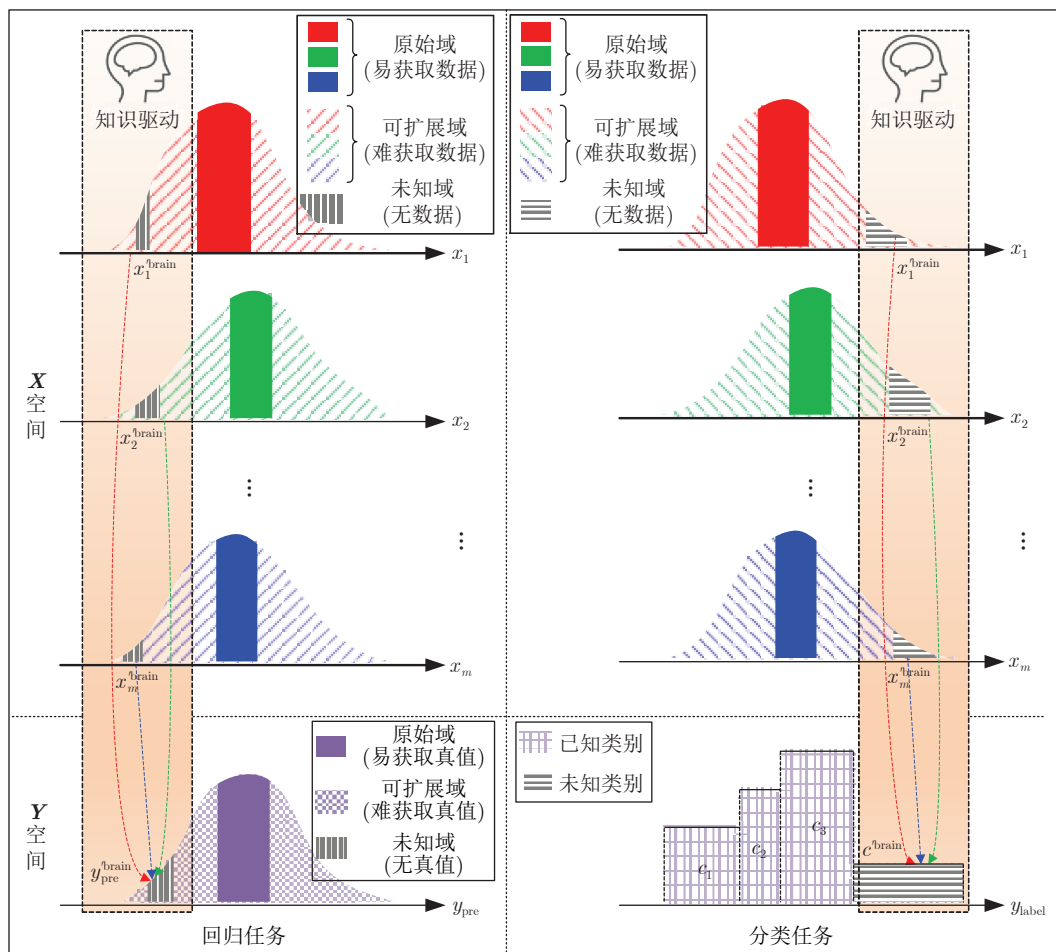


图 10 面向 VSG 的原始域、可扩展域和未知域的示意图

Fig. 10 Schematic diagram of original, extension, and unknown domain for VSG

此, Huang<sup>[94]</sup> 提出通过模糊数学进行样本集值化的处理方法, 即信息扩散, 其原理为: 通过三角、正态以及梯形等隶属度函数确定样本所蕴含信息的扩散范围. 在此基础上, Huang 等<sup>[95]</sup> 将正态扩散函数与神经网络相结合提出扩散神经网络 (Diffusion neural network, DNN), 但该方法仅适用于特征间的相关系数大于 0.9 的情况. 随后, Li 等<sup>[96]</sup> 在 DNN 的基础上提出大趋势扩散 (Mega-trend-diffusion, MTD) 技术, 如图 11 所示.

在图 11 中,  $m$  和  $n$  表示 2 个给定数据,  $b$  和  $a$  表示扩散函数的上界和下界,  $u_{set}$  表示样本变量取值的中心.

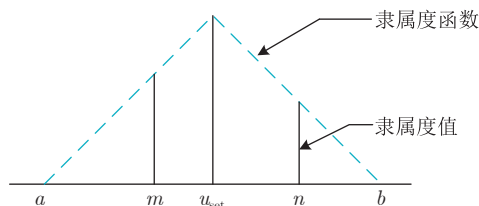


图 11 大趋势扩散技术

Fig. 11 Mega-trend-diffusion technology

由上可知, MTD 假设特征变量间相互独立和能够不对称地扩展特征范围, 进而能够在可扩展域上基于采样方式生成虚拟样本; 进一步, Lin 等<sup>[97]</sup> 提出广义趋势扩散 (Generalized-trend-diffusion, GTD) 技术, 即通过计算连续数据之间的趋势以获得序列数据的时间依赖性, 并采用所生成的虚拟样本解决柔性制造系统调度建模问题. 此外, Li 等<sup>[98]</sup> 通过集成 MTD 和树模型提出基于树结构趋势扩散 (Tree structure based trend diffusion, TTD) 的 VSG, 在多层陶瓷电容的介电系数预测实验中验证了其有效性. Rahimi 等<sup>[99]</sup> 提出基于神经网络的 MTD, 采用生成的虚拟样本构建聚合物 CO<sub>2</sub> 预测模型. 朱宝等<sup>[100]</sup> 提出采用三角分布和均匀分布共同表征小样本特性的多分布 MTD (Multi-distribution MTD, MD-MTD) 技术, 如图 12 所示.

在图 12 中, MD-MTD 采用三角分布在原始域样本空间表示真实小样本的分布情况, 采用均匀分布在可扩展域样本空间生成虚拟样本.

Sivakumar 等<sup>[101]</sup> 提出基于  $K$  近邻 ( $K$ -nearest neighbor, KNN) 的 MTD, 其通过原始样本的 KNN 计算扩展范围以确保虚拟样本的合理分布. Khamis 等<sup>[102]</sup> 提出基于  $K$ -means 的改进 MTD, 主要创新在于解决隶属度函数构建过程中的属性冗余问题. 此外, 也有研究人员采用组合多种信息扩散技术的策略生成混合虚拟样本, 如: 高克铨等<sup>[103]</sup> 提出

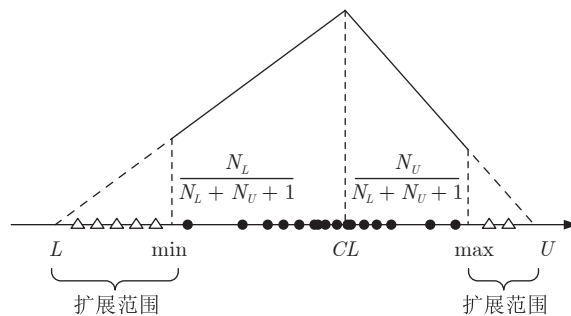


图 12 MD-MTD 示意图

Fig. 12 Schematic figure of multi-distribution MTD

改进型 MTD (Advanced MTD, AD-MTD), 结合文献 [100] 所提出的 MD-MTD 获得混合整体 MTD (Hybrid-MTD), 充分利用各自优势. 研究人员也提出结合 MTD 与其他插值方式的策略, 如乔俊飞等<sup>[56]</sup> 同时采用了 MTD 和隐含层插值.

与直接在可扩展域样本空间内以采样的方式获取虚拟样本不同, Li 等<sup>[46]</sup> 在采用 MTD 确定扩散范围后, 先基于遗传算法 (Genetic algorithm, GA) 生成虚拟样本输入, 再通过基于可行性的规划 (Feasibility-based programming, FBP) 模型生成虚拟样本输出; Chen 等<sup>[104]</sup> 先提出基于三角隶属度的信息扩散 (Information-expanded based on triangular membership, TMIE) 技术, 再在确定的范围后采用 PSO 算法获得虚拟样本输入, 最后通过 RWNN 得到虚拟样本输出. 此外, 针对不同 VSG 所产生虚拟样本间存在的冗余性与互补性, 汤健等<sup>[61]</sup> 采用 PSO 算法对基于领域专家知识和 MTD 生成的虚拟样本进行优化选择.

粗糙集理论是由 Pawlak 提出的处理具有模糊、不一致和不确定等特性数据的数学理论<sup>[105]</sup>, 其核心思想是从近似空间导出上近似算子和下近似算子 (又称上、下近似集), 将知识空间划分为上近似域、下近似域和边界域, 其中: 上近似域是由知识空间内与某一概念有非空交集的知识粒的并集构成的集合, 下近似域是由知识空间内包含某一概念的知识粒的并集构成的集合. 目前, 粗糙集理论已广泛应用于工业制造<sup>[106]</sup>、废水处理<sup>[107]</sup> 以及优化控制<sup>[108]</sup> 等领域, 但将粗糙集理论直接应用于回归 VSG 的研究还未见报道.

#### b) 分布假设

高斯分布是工业过程数据最为符合的假定分布. 文献 [109] 通过划分区间提出改善核密度估计 (Improved kernel density estimation, IKDE) 并生成虚拟样本以解决制造系统早期阶段样本少的问题; 随后, 文献 [110] 将 IKDE 扩展为通用模型, 应用于

具有时间依赖性的小样本建模问题并生成虚拟样本. Li 等<sup>[111]</sup>采用小型约翰变换方法 (Small Johnson data transformation, SJDT) 使得小样本数据趋近正态分布, 进而生成虚拟样本. 但是, 实际工业过程数据的期望分布不但未知且在小样本情况下也难以确定.

相较于高斯分布, 威布尔 (Weibull) 分布在工业制造、可靠性分析等领域的应用更为广泛. Li 等<sup>[112]</sup>针对产品寿命性能评估样本数量少的问题, 基于双参数 Weibull 分布选择最大  $p$  值 (Maximal  $p$  value, MPV) 的反直觉假设检验方法近似估计非线性和非对称的小样本分布, 并采用从分布中随机采样的策略生成虚拟样本, 但其实用性有待验证; 接着, Li 等<sup>[113]</sup>为解决 TFT-LCD 制作领域中的多模态小样本问题, 先用赤池信息准则 (Akaike information criterion, AIC) 的改进版 AICc (Corrected version of the AIC) 对聚类结果进行评价, 再通过 MPV 计算多峰分布的参数以确定虚拟样本数量, 最后生成虚拟样本, 但其适用性有待评估.

#### c) 基于知识

面向回归问题, 针对如图 10 所示的可扩展域和未知域, 可能存在不合理的虚拟样本和无法生成的虚拟样本; 此时, 需要借助工业过程自身机理知识和其他相似过程的经验知识予以辅助.

目前, 已有的基于领域知识的研究仅见于文献 [15], 其依据 MSWI 过程 DXN 值的下限范围进行可扩展域真值的修订. 如何借助工业过程的机理知识辅助回归问题 VSG 的研究还未见报道.

### 2) 面向分类问题的 VSG

#### a) 集合理论

文献 [114] 采用基于模糊的信息分解 (Fuzzy-based information decomposition, FID) 为少数类生成虚拟样本以平衡训练数据并对缺失值进行填充. Ramentol 等<sup>[115]</sup>基于 SMOTE 和粗糙集理论生成虚拟样本以处理不平衡数据集; 在此基础上, 胡峰等<sup>[116]</sup>提出基于三支决策的不平衡数据过采样策略, 首先依据样本总体分布定义正域、边界域和负域后, 再在边界域和负域生成虚拟样本, 结果表明可有效地解决不平衡数据的二分类问题, 但如何解决多分类问题仍有待研究. 由上可知, 基于集合理论面向分类问题 VSG 的研究还有待深入.

#### b) 分布假设

Yang 等<sup>[43]</sup>在假设过程数据符合高斯分布的基础上, 在计算其均值和标准差后采样生成虚拟样本, 实验表明采用适当数量的虚拟样本能够提高分类器泛化性能, 但如何确定数量未予以考虑; 进一步, Shen

等<sup>[117]</sup>在采用最大期望算法计算高斯模型的参数和采用 AIC 与贝叶斯信息准则 (Bayesian information criterion, BIC) 自适应确定模型高斯分量的最佳数量后, 通过采样获得虚拟样本. 文献 [118] 采用 SVM 的状态函数近似样本分布并通过采样输出虚拟样本. 文献 [119] 采用  $K$  均值聚类法检测多模态 Weibull 分布, 利用真实和虚拟样本间 Weibull 偏差的误差变化作为虚拟样本数量的评估标准.

#### c) 基于知识

面向分类问题, 在实际过程中存在无目标类别的样本用于模型训练的情况, 即零样本问题, 例如: 故障诊断领域存在特殊故障的样本无法获得的问题. 笔者认为, 类似于回归问题, 向生成过程添加机理或经验知识是解决 VSG 中未知领域零样本问题的有效手段.

实际工业过程中, 领域专家借助于对复杂机理的认知, 再辅以长期的工作实践和经验积累, 对已经出现的或可能出现的各种异常故障形成了相应的知识体系<sup>[120-121]</sup>. 研究表明, 将专家知识转换为属性、文本/语义、知识图谱、规则以及本体等融入到模型训练中, 可有效提高模型的泛化性和可解释性<sup>[122]</sup>. 对此, Link 等<sup>[121]</sup>采用基于专家知识定义的由故障位置、影响和原因等属性组成的故障描述确定故障类型, 相关的属性知识可从其他易获取的故障中预先学习和迁移, 故无需额外的训练数据. 但这种方法并无虚拟样本产生. Zhuo 等<sup>[123]</sup>提出基于故障属性 GAN (Fault attributes GAN, FAGAN) 的任意样本学习策略, 本质上是专家知识定义的故障属性作为辅助信息使得生成样本更接近真实样本, 实现对未知故障的诊断.

相较于领域专家直接提供的专家知识, 模型知识是通过对模型的学习和推导所提取出的隐含知识<sup>[124]</sup>. Yao 等<sup>[125]</sup>提出结合联邦学习和迁移学习的缺失数据填充策略, 目的是使不同边缘设备上的模型能够互相传递和利用所学习到的知识, 从而提高数据填充的准确性. Feng 等<sup>[126]</sup>提出基于多头语义表示和层次对齐技术的语义细化 WGAN (Semantic refinement WGAN, SRWGAN), 其通过细化粗粒度语义描述消除类别之间的偏差, 进而提高特征生成和知识转移的效果. 目前暂无基于模型知识驱动的工业过程 VSG 成果报道.

如何获得相关领域的专家知识和如何利用数值仿真模型提取符合工业过程的知识, 是未来支撑知识驱动 VSG 和解决未知域故障诊断的重要研究方向.

综上所述, 基于模糊集理论 VSG 的成果较为

丰富, 特点是: 面向回归问题的研究多于分类问题, 面向虚拟样本输入空间的研究多于输出空间. 此外, 目前的信息扩散技术缺少工业过程机理知识的支撑. 相较于模糊集理论, 粗糙集理论在 VSG 领域的研究较少, 所提知识空间的 3 个域并未给出相应的域扩展计算方法, 这将是未来基于粗糙集理论 VSG 的研究方向之一. 此外, 如何基于知识确定符合复杂工业过程的分布类型及相关参数是基于分布假设 VSG 的未来重要研究方向. 基于知识的 VSG 还处于辅助阶段, 相对而言在分类问题上更易研究.

## 2.2 基于 VSG 实现流程分类的研究现状

### 2.2.1 面向回归问题的 VSG 实现流程

#### 1) 过程数据预处理阶段

对过程数据进行预处理的目的是使得原始域样本空间的稀疏区域易于发现以降低 VSG 的难度. 首先进行对数据缺失值的处理, 如: 文献 [15] 和 [61] 对 MSWI 过程 DXN 数据中的缺失值进行删减和人工填充, 文献 [64] 和 [127] 对化工过程数据的异常和缺失值进行识别和去除, 文献 [125] 利用联邦学习和迁移学习进行缺失值填充. 然后, 采用特征工程进行数据处理, 如: 文献 [52] 和 [53] 采用 LLE 和 Isomap 从高维数据中提取 2 维特征, 文献 [54] 采用 *t*-SNE 提取 3 维特征, 文献 [59] 和 [128] 基于化工机理选择与运行指标相关的特征, 文献 [61] 基于专家经验选择与 DXN 排放浓度相关的特征. 最后, 进行标准化或归一化处理, 目的是消除不同特征上差异化数量级所造成的影响.

#### 2) 虚拟样本输入生成阶段

针对原始域样本空间而言, 通常是先采用某种方法识别原始域样本空间的稀疏区域后再采用各种策略生成虚拟样本输入, 如: 文献 [127] 采用欧氏距离识别稀疏区域后采用插值策略, 文献 [57] 通过对投影点的最大间距进行稀疏检测后采用中点插值; 文献 [52-54] 提取过程数据特征后通过可视化样本分布确定稀疏区域. 针对不同区域, 文献 [49] 提出稀疏假设和集中假设, 指出相较于在密集区域生成虚拟样本而言, 在稀疏区域生成虚拟样本更有必要, 但这需要权衡两个区域所生成的虚拟样本数量. 此外, 文献 [64] 和 [65] 通过 WGAN-GP 和 CWGAN 学习原始样本分布后生成虚拟样本输入.

针对扩展域样本空间而言, 采用信息扩散和分布假设等方法先确定可扩展区域再生成虚拟样本输入, 如: 文献 [96] 采用基于三角隶属度函数的 MTD 获得扩展范围后采用基于插值的生成策略; 文献 [104]

采用基于非对称三角隶属度函数的信息扩散技术获得扩展域范围, 通过 PSO 在该范围内生成虚拟样本输入; 文献 [129] 采用流形子空间对原始域真实样本进行分组并基于 MTD 确定扩展范围, 根据两者构建超球体方程后在球面和球内通过采样生成虚拟样本输入.

#### 3) 虚拟样本输出生成阶段

通常采用原始域的真实小样本训练的映射模型为虚拟样本输入匹配输出, 常用的映射模型包括 RWNN<sup>[59, 104, 129]</sup>、BPNN<sup>[52]</sup>、RF<sup>[54]</sup> 和 RBF<sup>[57]</sup> 等. 面向 GAN 策略, 文献 [64] 基于 CS-CGAN 和一致性检验为 WGAN-GP 所生成的虚拟样本输入匹配输出; 文献 [65] 将基于深度神经网络的回归器与生成器以及判别器共同训练以生成虚拟样本输出; 文献 [55] 采用分位数回归网络, 在一定置信度下为 CGAN 生成合适的虚拟样本输出, 进而减少生成器和判别器的训练难度.

#### 4) 虚拟样本质量筛选阶段

常见的虚拟样本质量筛选方法如下:

a) 相似性度量: Kullback-Leibler (KL) 散度<sup>[59]</sup> 和 Wasserstein 距离<sup>[64]</sup> 等方法因不能同时考虑输入和输出之间的关系而只能用于虚拟样本输入的度量, 不能直接用作回归问题中输入/输出虚拟样本对的筛选标准;

b) 优化算法: 文献 [61] 采用 PSO 算法对虚拟样本进行优化选择以提高样本质量;

c) 模型误差: 文献 [46] 指出合格虚拟样本构建模型的相对误差应小于 10%. 其他的相关研究包括: 文献 [130] 基于隶属度函数值的似然评估机制进行筛选, 文献 [49] 基于领域专家判断虚拟样本的合理性等. 综上, 笔者认为, 针对输入/输出虚拟样本对的筛选准则的研究还有待深入, 并且需要结合质量判别准则进行优化选择.

#### 5) 虚拟样本数量确定阶段

常用确定虚拟样本数量的方式是凭借经验或者依据逐批添加虚拟样本至真实小样本后所构建不同模型的泛化误差.

在此基础上, 文献 [131] 根据真实小样本的方差上限, 提出先采用信息熵理论确定虚拟样本数量再建立最优虚拟样本生成数量的概率模型的 2 步策略. 具体的, 确定虚拟样本数量的公式如下:

$$n_{\text{vir1}} \approx \sigma_0 \sqrt{2\pi e} - n_1 \quad (6)$$

式中,  $\sigma_0$  为真实样本的标准方差;  $n_1$  为原始样本的数量. 面向噪声 0.95 置信水平的最优虚拟样本概率模型的公式如下:

$$n_{\text{vir}2} = \left\lceil \frac{(2\mu_0 + 2.706) - \sqrt{(2\mu_0 + 2.706)^2 - 4\mu_0^2 C_0}}{2\mu_0^2} \right\rceil \quad (7)$$

式中,  $\mu_0$  为真实小样本的均值,  $C_0$  为虚拟样本产生的总噪声.

文献 [49] 根据其所提的稀疏假设和集中假设, 给出如下的虚拟样本数量确定公式:

$$n_v = 10^{\lceil 1 + \log_{10} n \rceil} \quad (8)$$

式中,  $n$  为训练样本数量;  $n_v$  为添加虚拟样本数量.

#### 6) 特殊阶段

目前, 已有研究学者提出, 先生成虚拟样本输出再匹配生成虚拟样本输入的“反向”VSG策略, 如: 文献 [20] 利用 LOF 获得原始样本输出的稀疏区域并通过  $K$ -means++ 获得中心点, 利用中点插值生成虚拟样本输出后将其作为 CGAN 的条件信息以生成虚拟样本输入; 文献 [49] 在获得原始样本输出的密集和稀疏区域并利用三样条插值生成虚拟样本输出后, 基于 ITNN 生成虚拟样本输入.

综上可知, 采用不同策略的 VSG 具有差异化的特性. 如何面向特定应用领域进行选择和改进是应用时需关注的问题.

### 2.2.2 面向分类问题的 VSG 实现流程

针对不存在未知类的分类问题而言, VSG 将类别信息直接作为先验用于虚拟样本输入的生成. 因此, 本节将从过程数据预处理、虚拟样本输入生成、虚拟样本质量筛选和虚拟样本数量确定等阶段进行综述.

#### 1) 过程数据预处理

由于采用 VSG 技术的故障数据多为机械信号, 常用方法是采用 FFT 将时域数据转换为频域数据. 也有学者将一维时域信号转换为二维图像后进行处理, 如: 文献 [85] 和 [132] 将振动信号切为若干片段, 依次归一化并取整后转换为灰度图进行 VSG. 研究学者提出根据计算机视觉领域的增强策略对机械信号进行处理以缓解生成模型过拟合现象的策略, 如: 文献 [133] 采用重叠分割、旋转和抖动的数据增强策略对故障样本进行处理, 文献 [92] 对故障样本进行平移和缩放处理等.

#### 2) 虚拟样本输入生成

SMOTE 作为针对少数类样本进行随机线性插值的 VSG 技术, 已广泛应用于解决类不平衡问题, 如: 文献 [115] 结合粗糙集理论生成少数类虚拟样本, 文献 [73] 对电力变压器非正常状态的样本进行补充, 文献 [74] 在支持向量近似的分类边界生成非正常状态的虚拟样本, 文献 [72] 采用 Minkowski 距

离代替传统 SMOTE 中的欧氏距离, 文献 [75] 通过控制生成范围减少重叠样本等.

随着深度学习的发展和 GAN 的提出, 对抗模型已成为面向分类问题 VSG 的研究热点. 为保证训练过程的稳定性和虚拟样本的质量, 目前研究主要集中在改进损失函数和模型结构. 改进损失函数的研究成果包括: 采用 Wasserstein 距离替换传统交叉熵损失函数的 WGAN<sup>[84]</sup>, 在 WGAN 损失函数的基础上增加梯度惩罚项的 WGAN-GP<sup>[85]</sup>, 采用 Pull-away 损失函数的改进 GAN<sup>[69]</sup> 等. 改进模型结构的成果包括: 文献 [79] 在 ACGAN 中添加 Dropout 层以缓解虚拟样本生成过程中的模式崩溃问题, 文献 [68] 采用 CVAE 取代 ACGAN 的生成器, 文献 [90] 采用 VAE 作为 GAN 的生成器并将遗憾算法用于判别器, 文献 [83] 采用包含具有真假判断、故障诊断和故障分类 3 种功能的判别器, 文献 [92] 采用 CVAE 作为 WGAN 的生成器并通过自调制算法进行更新以提高模型稳定性, 文献 [19] 采用并行 GAN 生成虚拟样本等. 上述这些研究, 在如何扩展虚拟样本输入的边界方面的研究较少, 原因在于分类问题在输出空间上相对于回归问题的特殊性.

此外, 受限于原始域样本空间所蕴含的机理知识匮乏的问题, 部分学者采用迁移学习从相似领域中提取知识用以辅助生成虚拟样本. 基于样本进行迁移的研究包括: Zhang 等<sup>[134]</sup> 提出结合 SMOTE 和迁移学习的 VSG 以处理不平衡数据, 采用源域样本和目标域原始样本加权的方式生成虚拟样本, 结果表明能够有效提高分类器的准确性; Liu 等<sup>[135]</sup> 采用自适应混合方法生成虚拟样本, 包括基于迁移学习策略保证生成样本的数量与多样性和通过进化算法提高故障诊断精度; 贾欣等<sup>[136]</sup> 提出将多数类样本迁移为少数类边界样本的均衡方案, 这有利于学习类别决策边界. 从模型角度进行迁移以生成虚拟样本的研究包括: 廖一帆等<sup>[137]</sup> 通过 Fine-tuning 方法将由临界和非临界样本所训练的预测模型嵌入 WGAN 中以辅助生成非临界样本; 兰健等<sup>[138]</sup> 通过 GAN 学习电力系统各种运行方式的共性特征, 之后基于微调得到高性能的典型运行方式生成模型, 为运行方式的分析提供支撑. 由上可知, 目前迁移学习已经成为 VSG 的研究热点之一, 但迁移后的可靠性方面还有待验证.

#### 3) 虚拟样本质量筛选

目前通常采用组合多种评价指标的方式对虚拟样本质量进行评估, 如表 1 所示.

由表 1 可知, 用于分类问题 VSG 的评价指标包括: Wasserstein 距离、欧氏距离、马氏距离、KL

表 1 面向分类问题的虚拟样本评价指标  
Table 1 Virtual sample evaluation index for classification problem

评价指标	文献	年份
Wasserstein 距离	[139]	2020
KL 散度、F-score、Kappa 系数、GAN 测试值	[66]	2021
Wasserstein 距离、KL 散度、欧氏距离、皮尔逊相关系数	[67]	2021
马氏距离、欧氏距离	[82]	2021
判别概率、最大均值差异、KL 散度	[69]	2022
皮尔逊相关系数	[85]	2022
最大均值差异、KL 散度、GAN 测试值	[92]	2022

散度、F-score、Kappa 系数、皮尔逊相关系数、判别概率、最大均值差异和 GAN 测试值等, 这表明目前还不存在统一标准, 相关的理论支撑也未见报道. 相对而言, 文献 [82] 和 [69] 给出了评价指标的具体阈值并据此进行样本筛选.

#### 4) 虚拟样本数量确定

面向分类问题 (以存在 A 和 B 两类为例, 其中 A 类数量远大于 B 类数量), VSG 的目的是: 通过生成 B 类虚拟样本降低数据集的不平衡比 IR 直至其值为 1, 即实现  $N_{\text{majority}}^A = N_{\text{minority}}^B$ . 从上述视角, 虚拟样本的理想数量即为 A (多数) 类和 B (少数) 类样本的数量之差, 可表示为:

$$N_{\text{vir}} = N_{\text{majority}}^A - N_{\text{minority}}^B \quad (9)$$

其中,  $N_{\text{vir}}$  为虚拟样本的数量,  $N_{\text{majority}}^A$  和  $N_{\text{minority}}^B$  分别为 A (多数) 类和 B (少数) 类样本的数量.

文献 [140] 指出, 生成虚拟样本并不需要完全地消除少数类与多数类之间的数量差距, 可通过类别之间的分类复杂度  $H_{\text{class}}$  确定最终所需虚拟样本数量, 如下:

$$N_{\text{vir}}^{\text{final}} = H_{\text{class}} \times N_{\text{vir}} \quad (10)$$

综上所述, 面向工业过程的 VSG 需要根据具体任务和实际数据的特性设计相应的 VSG 流程和采用适合策略.

## 2.3 基于 VSG 推广应用分类的研究现状

本文依据当前工业过程中 VSG 的研究现状, 从回归和分类两类问题对 VSG 的推广应用情况进行综述.

### 2.3.1 面向回归问题的 VSG 应用

目前 VSG 主要应用于石油化工、固废焚烧、工业制造和矿业冶金等领域, 其统计结果如图 13 所示.

如图 13 所示, VSG 在工业制造和石油化工领域应用和发展时间较长, 而在固废焚烧和矿业冶金领域的应用才刚刚起步.

面向化工过程, 文献 [104] 提出基于信息扩散和 PSO 优化的 VSG, 通过 RWNN 为虚拟样本输入匹配输出, 提高了精对苯二甲酸 (Pure terephthalic acid, PTA) 生产过程的醋酸消耗预测模型的性能; 文献 [59] 通过在 RWNN 隐含层间插值生成虚拟样本输出和虚拟样本输入, 构建乙烯生产系统模型以为石化行业的能源管理提供指导作用; 文献 [53] 针对 PTA 生产过程的数据分布不完备问题, 采用 Isomap 流形学习进行降维并搜寻稀疏区域插值生成虚拟样本, 结果表明该方法可有效提高软测量模型的性能; 文献 [55] 将分位数回归神经网络嵌入至 CGAN 内为虚拟样本匹配准确输出, 采用实际过程数据验证了所提方法的有效性; 文献 [65] 将回归器嵌入至 CWGAN 中, 针对 PTA 生产过程的应用表明, 所生成的虚拟样本质量优于常规方法.

针对 MSWI 过程 DXN 建模数据获取困难的

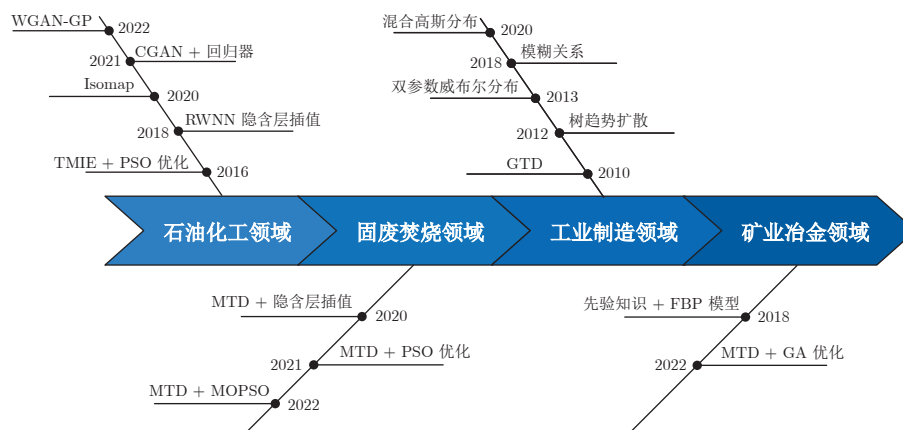


图 13 面向回归建模问题的 VSG 应用统计结果

Fig. 13 VSG application statistical results for regression modeling problem

问题, 文献 [56] 提出基于改进大趋势扩散和隐含层插值的混合 VSG, 即生成的虚拟样本包含基于子区域欧氏距离改进的 MTD 等间隔生成和基于正则化改进的 RWNN 隐含层插值生成两类, 通过混合样本构建 DXN 排放软测量模型, 但样本间的冗余性未予考虑; 接着, 文献 [61] 基于领域专家知识和 MTD 技术对真实样本进行扩展, 在生成虚拟样本输入和根据映射模型获得虚拟样本输出后, 采用 PSO 算法对虚拟样本进行优化选择, 但该方法未能同时考虑虚拟样本数量和映射模型超参数对模型泛化性能的影响; 对此, 文献 [15] 提出基于多目标 PSO (Multi-objective PSO, MOPSO) 混合优化的 VSG, 通过对虚拟样本数量和模型性能指标 2 个目标进行混合优化的策略确保了 VSG 的合理性和有效性。

在工业制造领域中, 文献 [97] 针对柔性制造调度系统建模过程中存在的样本信息匮乏且与时间相关的问题, 提出基于 GTD 技术的 VSG, 结果表明混合样本有助于提高模型性能; 在此基础上, 文献 [98] 将趋势扩散和树算法结合, 提出基于树结构的趋势扩散方法, 用于扩充制造过程初期的样本数量; 针对产品寿命性能评估问题, 文献 [112] 在符合制造业的 Weibull 分布中, 以采样方式获得虚拟样本; 文献 [141] 采用模糊  $c$  均值聚类算法将数据分为多个簇后赋予不同权重, 通过箱型图估计特征的扩展范围后生成虚拟样本, 构建的模型相较于对比方法具有更佳的性能; 文献 [117] 采用高斯混合模型拟合数据分布后采用网格搜索技术对模型进行优化, 所提方法能够缓解橡胶加工耐磨性数据的缺乏和提高预测模型的精度。

针对磨矿过程采用非完备样本构建数据驱动模型困难的问题, 文献 [6] 提出结合先验知识和 FBP 的 VSG, 对构建物理阐释明确的软测量模型具有重要的借鉴意义。针对稀土萃取过程中存在的小样本

问题, 文献 [142] 将基于 MD-MTD 和 RWNN 生成的虚拟样本与 GA 优化 MD-MTD 生成的虚拟样本混合后构建预测模型, 结果表明可提高模型的稳定性和泛化性能。

针对其他领域回归问题的 VSG 还包括: 锂电池剩余寿命预测<sup>[143]</sup>、蒸馏塔煤油凝固点预测<sup>[144]</sup>和血液光谱分析<sup>[103]</sup>等。

### 2.3.2 面向分类问题的 VSG 应用

已有研究成果主要集中在故障诊断领域, 即用于轴承、齿轮以及电机等机械设备诊断模型的故障样本生成。与传统过程数据不同, 这类故障样本多为用于二分类问题的机械振动信号, 特点是故障样本的数量明显少于健康样本, 即存在类不平衡问题<sup>[145]</sup>。基于这一特点, 图 14 给出了 2019 ~ 2022 年间 VSG 技术在故障诊断领域的应用。

由图 14 可知, 近 4 年故障诊断领域的 VSG 研究成果主要集中于编码器、GAN 等深度学习方法, 其本质是通过改进生成模型的结构、数量和损失函数等方式保证虚拟样本的质量。

在滚动轴承故障诊断领域中, 文献 [146] 提出结合迁移学习和 GAN 的 VSG, 其基于设备故障机理对特征进行迁移并通过 GAN 学习设备监测数据的分布特征进而生成虚拟样本, 具有较好的变工况迁移能力; 文献 [147] 提出融合用于生成虚拟故障样本的去噪自编码器 (Predictive generative denoising AE, PGDAE) 和进行故障诊断的深度珊瑚网络 (Deep coral network, DCN) 的统一框架, 结果表明可有效生成虚拟故障样本并准确识别滚动轴承故障; 为提高生成模型性能, 文献 [66] 采用元学习增强 Wasserstein AE (WAE) 策略提升先验分布与滚动轴承振动信号间的映射能力, 结果表明生成的虚拟样本质量优于对比方法; 文献 [92] 将自调制、CVAE 和 WGAN 相结合以增强博弈对抗过程的

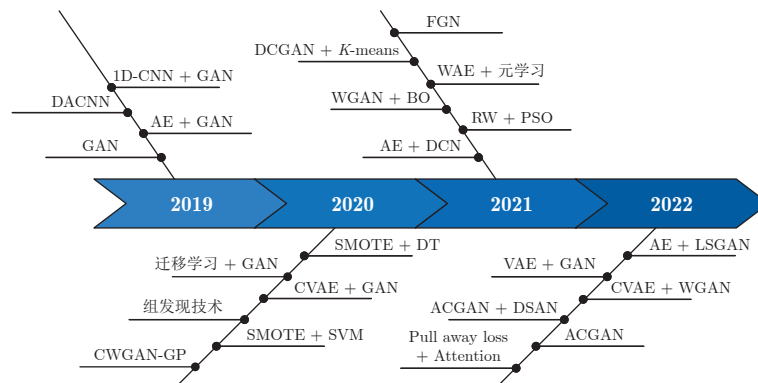


图 14 2019 ~ 2022 年面向故障诊断领域的 VSG 应用统计结果

Fig. 14 VSG application statistical results for fault diagnosis on 2019 ~ 2022

稳定性, 进而生成高质量的虚拟故障样本; 文献 [132] 将一维信号数据转换成二维灰度图像后基于 ACGAN 生成虚拟故障样本, 通过自注意力机制深度子域适应网络 (Deep subdomain adaptation network, DSAN) 提高故障特征的非线性拟合能力; 文献 [148] 利用常数  $Q$  转换将机械信号转为频谱图像后输入 GAN, 并采用均方误差替换交叉熵作为损失函数; 文献 [149] 通过度量判别器与生成器间的相对性能后自适应调节生成器的损失值, 结果表明对抗学习过程的收敛更快, 所生成样本的质量更好; 文献 [69] 提出深度特征增强生成对抗网络以提高不平衡故障诊断的性能, 建立自动数据过滤器以保证生成样本的准确性和多样性; 文献 [150] 提出深度特征生成网络 (Deep feature generating network, DFGN) 用于面向零样本的滚动轴承故障检测, 实验结果表明能够有效地检测典型故障。

面向变压器的故障样本与健康样本不平衡的问题, 文献 [73] 采用 SMOTE 生成异常状态样本后, 通过 DT 提取变压器状态评估知识后将其转换为状态量和评估规则; 文献 [74] 在支持向量近似的分类边界, 根据最近邻决策机制采用插值方式生成虚拟样本, 进而提高诊断模型的准确性; 文献 [151] 采用基于梯度惩罚优化的 CWGAN (CWGAN-GP) 生成多类别故障样本, 构建基于栈式自编码器的诊断模型, 结果表明可有效改善模型分类偏好和提升分类性能; 文献 [91] 设计包含孪生编码器、解码器和传输层的改进 AE 以消除数据噪声, 通过 LSGAN 生成高置信度的健康状态样本, 结果表明能够及时检测发电机的潜在异常情况。

针对风力涡轮机故障样本稀少引起的信息缺失问题, 文献 [152] 按照皮尔逊相关系数和最大信息系数, 将生成的虚拟样本特征分组输入判别器后分别计算损失, 以加权值作为总损失用于更新 GAN, 实验表明所生成的虚拟样本更为真实; 文献 [153] 提出将对抗学习作为正则项引入 CNN 的深度对抗 CNN (Deep adversarial CNN, DACNN), 结果表明提高了诊断模型的准确度; 文献 [89] 针对机械系统的异常样本采集难的问题, 提出结合 GAN 和 AE 的机械系统异常检测方法, 通过编码-解码-再编码的网络结构学习异常变化并生成虚拟样本, 结果表明能够更稳定地表征故障演化过程; 文献 [154] 提出将样本生成和故障诊断相结合的 ASM1D-GAN (Assembled 1D CNN and GAN), 通过对抗学习机制同时优化上述两个过程以达到同时提高生成样本质量和故障诊断精度的目的。

针对齿轮箱的故障诊断问题, 文献 [68] 提出基

于条件变分自编码器生成对抗网络的不平衡故障诊断方法, 通过 CVAE 提取故障样本分布以对抗方式生成虚拟样本, 结果表明可生成不同工况下的故障样本, 能够提高模型性能; 针对 GAN 调参复杂且具有随机性的问题, 文献 [86] 通过贝叶斯优化 (Bayesian optimization, BO) 策略自适应地调节 WGAN 的判别器参数以提升虚拟样本质量, 结果表明可有效提高故障识别精度; 文献 [67] 采用深度卷积 GAN (Deep convolution GAN, DCGAN) 生成虚拟样本以解决数据集不平衡问题, 通过  $K$ -means 聚类算法改进基于 CNN 的机械设备故障诊断模型。

此外, VSG 在故障诊断中的应用还包括: 小电流接地系统故障线路检测<sup>[77]</sup> 以及热电联产电厂给水泵<sup>[155]</sup>、磨矿机<sup>[156]</sup> 和化工过程<sup>[157]</sup> 等领域的故障诊断。

综合上述研究可知, VSG 正快速在缺失完备建模样本的复杂工业过程中获得应用, 其在面向分类问题的研究深度和先进性等方面明显优于面向回归问题。本文虽然仅对常见的工业过程的 VSG 典型应用进行了介绍, 但这些结果在一定程度上表明, VSG 具有独特的优势和适应不同工业过程数据的良好性能。

### 3 数据集与开源软件

本节将对上述面向工业过程的 VSG 研究所涉及的数据集和开源软件进行总结, 包括用于虚拟样本实验评估的基准数据集和在 VSG 算法实现过程中所用到的开源软件, 进而为 VSG 的研究提供基础支撑。

#### 3.1 基准数据集

本节将从面向回归和分类问题两个方面对目前 VSG 研究中采用的合成和公开基准数据集进行汇总, 如表 2 和表 3 所示。

由表 2 和表 3 可知, 目前的 VSG 研究大多是在传统的合成函数和公开的故障诊断数据集上开展的, 基于实际工业过程的 VSG 基准数据集还未见相关报道, 尤其面向回归问题, 甚至不存在由实际工业过程产生的数据集。因此, 构建能够用于生成模型和虚拟样本质量评估的通用 VSG 基准工业数据集也是未来的重要研究方向之一。构建面向实际工业过程的虚拟样本数据库更是值得深入研究的基础工作。

#### 3.2 开源软件

合适的编程软件是实现 VSG 的重要基础, 目

表 2 面向回归问题 VSG 的合成数据集  
Table 2 Synthetic datasets of VSG for regression problem

基准函数	取值空间	文献
$y = \begin{cases} \sin x/x, & \text{if } x \neq 0 \\ 1, & \text{if } x = 0 \end{cases}$	$x \in [-2\pi, 2\pi]$	[49]
$y = 2.077\ 5 + 9.045\ 46 \times (10^{-1}) x_1 + x_2^2 + \cos(x_3) + 1.355\ 6 \times (1.5 \times (1 - x_4)) + x_5^3 + x_6 - 2.571\ 51x_7 - 5.097\ 36 \times (10^{-1}) \times (x_8^2)$	$x \in [0, 1]$	[53]
$y = 0.415 \sin x_1 - 0.312x_2^2 + 1/(1 + e^{-x_3}) + \cos x_4^3 + 0.66e^{1-x_5^{0.5}} \sin x_5 - \cos x_6 \ln(1/\cos x_6) + 0.38 \tanh x_7 + (1 - x_8^3) \cos x_8^3$	$x \in [0, 1]$	[54]
$y = x + \varepsilon, \varepsilon \sim N(0, 0.05^2)$		
$y = x + \varepsilon, \varepsilon \sim N(0, 0.01x^2)$	$x \in [0, 1]$	[55]
$y = x + 0.2 \sin(20x) + \varepsilon, \varepsilon \sim N(0, 0.05^2)$		
$y = 1.335\ 6 \times (1.5(1 - x_1)) + e^{2x_1-1} \sin(3\pi(x_1 - 0.6)^2) + e^{3(x_2-0.5)} \sin(4\pi(x_2 - 0.9)^2)$	$x \in [0, 1]$	[57, 64, 127]
$y = \sin(x_1) + \cos(x_2) + \sin(x_1) \times \cos(x_2)$	$x_1 \in [-\pi, \pi]$ $x_2 \in [0, 2\pi]$	[59, 129]
$y = e^{(2x-1)} \sin[4\pi(x - 0.6)^2] + \varepsilon, \varepsilon \sim N(0, 0.002\ 5)$	$x \in [0, 1]$	[20]

注:  $\varepsilon$  是为了更好地模拟实际工业过程的环境影响而添加的噪声项。

表 3 面向分类问题的 VSG 公开数据集  
Table 3 Public datasets of VSG for classification problem

数据集	数据集信息	文献
Case Western Reserve University (CWRU) 轴承故障数据集 <sup>1</sup>	由美国凯斯西储大学发布的位于轴承数据中心网站的轴承故障数据集, 包含无故障和滚动体、内圈和外圈故障数据	[67, 69, 92, 132-133]
University of Connecticut (UoC) 齿轮箱故障数据集 <sup>2</sup>	美国康涅狄格大学 Jiong Tang 团队发布的齿轮箱故障数据集, 包括健康工况、缺齿、齿根裂纹、齿面剥落以及不同程度齿尖破损状态数据	[85]
Tennessee Eastman process (TEP) 数据集 <sup>3</sup>	由美国伊士曼化学公司开发的化学过程模拟平台生成的数据集, 包括正常工况和 21 种异常工况数据	[156-157]
IEEE PHM 2009 齿轮箱故障数据集 <sup>4</sup>	由 2009 年的 IEEE PHM 挑战赛提供的齿轮箱故障数据集, 包含健康、缺齿、齿裂等 8 种工况	[80]
西安交通大学 Spectra Quest (SQ) 轴承故障数据集 <sup>5</sup>	由西安交通大学 SQ 实验平台得到的电机轴承外圈和内圈故障数据集	[150]

数据集网址:

<sup>1</sup> <https://engineering.case.edu/bearingdatacenter/download-data-file>

<sup>2</sup> [https://figshare.com/articles/dataset/Gear\\_Fault\\_Data/6127874/1](https://figshare.com/articles/dataset/Gear_Fault_Data/6127874/1)

<sup>3</sup> <http://depts.washington.edu/control/LARRY/TE/download.html>

<sup>4</sup> <http://www.phmsociety.org/references/datasets>

<sup>5</sup> <https://github.com/sliu7102/SQ-dataset-with-variable-speed-for-fault-diagnosis>

前主要分为 Python 和 Matlab 两类。

#### 1) 基于 Python 的开源软件

a) PyTorch, 由 Facebook AI Research 开发的深度学习库, 支持基于 CPU 和 GPU 进行高效张量运算并提供可灵活修改模型结构的动态计算图, 包

含许多深度学习模型和算法, 详见官网: <https://pytorch.org/>.

b) TensorFlow, 由 Google Brain 团队开发的机器学习平台, 支持 GPU 和 TPU 等硬件加速计算并能够进行分布式的训练和推理, 提供了丰富的工

具和资源。除 Python 外, TensorFlow 还支持 Java、C++ 等编程语言, 详见官网: <https://www.tensorflow.org/>.

c) Keras, 由 Python 编写的开源神经网络库, 能够在 TensorFlow、CNTK 以及 Theano 上运行, 支持快速实验和构建复杂模型, 详见官网: <https://keras.io/>.

#### 2) Matlab 的开源软件

a) Deep Learn Toolbox, 其包含多种模型、算法和应用程序的深度学习框架, 支持网络设计可视化和训练进度实时监控, 详见官网: <https://ww2.mathworks.cn/products/deep-learning.html>.

b) Statistics and Machine Learning Toolbox, 其提供多种用于数据描述、分析和建模的有监督、半监督和无监督机器学习算法, 能够自动生成 C/C++ 代码用于嵌入式部署, 详见官网: <https://ww2.mathworks.cn/products/statistics.html>.

目前, VSG 研究正处于与统计学习、深度学习、迁移学习、联邦学习、集成学习等领域的新进展深度结合阶段, 因此这些领域所采用的开源软件都可用于 VSG 领域。进一步, 后续研究可考虑构建由基础算法、基准数据集、标准评估算法以及可视化等组件组成的 VSG Toolbox。

## 4 VSG 的比较与讨论

### 4.1 方法比较

从样本覆盖区域、VSG 实现流程和推广应用 3 个方面, 针对回归问题和分类问题 VSG 的研究成果统计与对比如附录 A 的表 A1 所示。文中的符号说明如表 A2 所示。

由表 A1 可知: 从 3 个不同视角综述的结果而言, 面向回归和分类问题的 VSG 在侧重点上是存在差异性的, 具体表现为:

1) 样本覆盖区域视角。面向原始域样本空间的 VSG 最早源于 SMOTE 等插值算法, 在 GAN 出现后其迎来了更高的研究热度, 其中: 分类问题主要集中在故障诊断领域, 采用卷积网络、编码器和注意力机制对故障数据进行特征提取和增强; 回归问题采用流形学习、专家经验等处理高维过程数据; 此外, 由于在博弈对抗的过程中为虚拟样本输入匹配准确的输出存在困难, 使得基于 GAN 的回归问题 VSG 研究较少。从通过 VSG 完备样本分布的目的的视角, 识别真实样本的稀疏区域是基于函数模型进行 VSG 的关键, 即首先通过稀疏区域确定需要生成虚拟样本的位置; 但是, 在基于对抗模

型的 VSG 过程中, 可能会生成不属于完备域 (期望域) 样本空间的不合格虚拟样本, 因此进行样本筛选很有必要。面向扩展域样本空间的 VSG 最早源于信息扩散理论, 在 MTD 提出后获得广泛应用, 其中: 基于模糊集理论的 VSG 研究相较于粗糙集理论更加成熟; 基于分布假设的 VSG 针对不同工业过程需选择合适的分布以接近完备域的样本分布。从完备样本分布区域的目的而言, 基于扩展域空间的 VSG 既要考虑可扩展区域存在与否和存在时的区域范围, 又要考虑扩展时虚拟样本的分布程度; 同时, 还需要结合知识对未知域进行认知。因此, 已有研究通常为原始域和扩展域分别选择合适的 VSG 策略。

2) VSG 实现流程视角。数据预处理阶段需要依据建模数据特性进行处理以便更好地开展后续工作, 例如: 对异常和缺失值进行剔除和填充, 对高维数据进行特征约简以及对机械信号进行 FFT 处理等。针对分类问题, 由于无需匹配输出, 采用 GAN 在虚拟样本输入生成阶段的研究明显多于回归问题, 这也导致基于扩展域样本空间的 VSG 研究较少。在虚拟样本输出生成阶段, 采用 RF、RWNN 和 RBF 等映射模型均能够适应小样本建模, 但如何基于有限的样本构建准确且鲁棒的映射模型仍是待解决的热点研究问题。在虚拟样本质量筛选阶段, 通常采用的是相似性度量、优化算法和模型误差等方法, 但如何确定统一的、有理论支撑的期望评价准则仍是一个未解决的开放性问题。在虚拟样本数量确定阶段, 目前多依据实际问题特性采用试凑方式确定添加数量, 虽有学者从数学理论和数据特性等角度探索确定方法, 但仍待继续完善。

3) VSG 推广应用视角。面向回归问题的 VSG 主要应用在石油化工、固废焚烧、工业制造和矿业冶金等领域, 其中: 石油化工领域多采用基于原始域样本空间的 VSG; 工业制造领域的 VSG 研究多集中于扩展域样本空间; 固废焚烧和矿业冶金的 VSG 研究相对较少, 处于起步阶段。面向分类问题的 VSG 应用在轴承、齿轮、涡轮机以及变压器等机械或电力设备的故障诊断中, 其中以面向机械信号采用 GAN 的应用最为广泛。

综上所述, 在上述 3 个视角下, 针对回归问题和分类问题的 VSG 各具优势, 有必要相互进行借鉴; 同时, 也有待于与迁移学习、集成学习、联邦学习等算法结合并与具体应用领域进行深度融合。

### 4.2 讨论与分析

结合以上分析, 笔者总结了面向工业过程 VSG 的未来研究方向, 如下所示。

### 1) 样本质量与生成模型协同优化

由式 (1) 可知, 采用增加样本数量或减少特征维数均是获得较大  $\alpha$  值的可行方案. 在 VSG 前基于特征工程降维以减少模型的训练难度是必要的, 其中: 基于特征变换 VSG 的难点在于如何重构虚拟样本, 基于特征选择 VSG 的难点在于如何平衡选择的特征数量和生成的虚拟样本质量等问题. 虚拟样本输出的质量在很大程度上取决于生成模型的选择, 但目前尚无统一的评估方式以分析模型结构或参数对虚拟样本的影响. 针对某个工业过程的某个实际问题所设计的 VSG 效果好但具有局限性, 如何借鉴并提高普适性有待研究. 因此, 设计虚拟样本质量评价指标并与生成模型的结构和参数协同优化是未来的重要研究方向, 同时也需要考虑如何提高优化效率、降低运行消耗等问题.

### 2) 基于对抗学习融合机理知识、经验规则和数据驱动模型的智能 VSG

现有 VSG 主要利用原始真实样本构建基于数据驱动的生成模型, 存在蕴含机理知识缺乏和完备样本分布未知等问题. 针对具体复杂工业过程而言, 可利用数值仿真软件构建能够反映运行状态的近似机理可视化模型和利用专家经验知识构建反映运行规则的经验模型. 因此, 通过对抗学习等技术自行选择由机理知识、经验规则和数据驱动等构成的多类型生成模型并通过进化最优 VSG 流程, 将能够为生成模型的选择和构建提供指导作用和提升 VSG 的可解释性.

### 3) 基于合成数据集的 VSG 理论分析

虽然 VSG 已在复杂工业过程的各个领域得到迅速发展, 但与其相关的理论分析却较为匮乏, 例如: 扩展域样本空间的隶属度函数和分布函数的选择依赖于主观经验; 用于信息扩散的三角隶属度函数和用于分布假设的正态分布函数并不适用于所有工业过程. 在优化算法领域中, 常采用多种基准函数进行算法设计、性能测试和方法比较, 依据这些人为设定的基准函数能够较为客观地评价不同优化算法的各种性能. 对此, 也有学者设计测试函数并采样得到合成数据对 VSG 性能进行评价<sup>[57, 129]</sup>. 但是, 在如何确定完备分布, 如何确定不同分布下虚拟样本的数量和质量等方面的理论还缺失. 因此, 笔者认为, 采用具有较好规范性和多样性的合成数据进行 VSG 的理论分析是未来该领域偏向于学术方面的研究方向之一.

### 4) 借鉴相关领域知识的迁移 VSG

不管是基于原始域样本空间还是基于扩展域样

本空间的 VSG, 本质都是基于原始真实样本并从中挖掘样本间的联系或获取扩展范围, 但受限于样本数量该过程存在多种困难. 以 GAN 为例, 其作为一种本身需要数据支撑的神经网络, 只有在存在充足数据时才能支持网络训练的收敛, 在数据量较少的情况下难以达到纳什均衡且易陷入模式崩塌, 此时的样本生成过程近似于对原始样本的简单复制, 显然这对提高样本的多样性和进行区域空间扩充并无实质性的帮助<sup>[158]</sup>. 因此, 除机理知识外, 从外部的样本空间获取知识以提高生成模型的性能是 VSG 的重要研究方向. 显然, 这种外部的样本空间应与原始域空间具有相似性且数据量大, 此处将其称为相似域空间. 迁移学习旨在利用相关领域的知识提高学习性能或最小化目标领域所需的样本数<sup>[159]</sup>. 目前, 基于相似域空间的 VSG 尚处于起步阶段, 还存在大量问题亟待解决, 例如: 两个域之间存在相似性是知识迁移的必要前提, 但相似性度量方法的优劣还未有统一标准; 域间相似性对虚拟样本质量的影响程度也是值得研究的问题; 如何从数据和模型两个层面同时进行迁移以达到更好的效果等.

### 5) 工业过程数字孪生系统驱动的 VSG 完备样本分布研究

工业过程数据存在样本稀缺、分布完备性差和内涵机理知识匮乏等问题. 如何获取具有完备样本分布的建模数据是未来 VSG 实现落地应用的关键. 近些年, 数字孪生技术的出现以及其迅速的发展为解决上述问题提供了新的思路. 文献<sup>[160]</sup>构建航天器电源系统的数字孪生模型, 并对其注入虚拟故障以获得虚拟样本. 文献<sup>[161]</sup>通过采煤机摇臂机的数字孪生模型生成状态检测样本并构建预测模型, 为复杂矿用设备的运维提供支持. 虽然上述数字孪生系统多面向离散过程, 但也能够为构建机理更加复杂的流程工业数字孪生系统提供借鉴. 因此, 基于物理几何模型和动力学模型以及多源数据构建复杂工业过程数字孪生模型, 生成具有完备样本分布的虚拟样本库能够为 VSG 提供机理知识, 具体实现方式与可用性验证等问题还有待研究.

### 6) 基于监督和半监督学习的集成 VSG

复杂工业过程的关键运行指标数量受限于检测技术的高成本和大时滞特性, 导致存在大量未标记的过程数据和少量标记的建模数据共存的现象<sup>[162]</sup>. 半监督学习是综合有标记和无标记数据的建模方法, 其能充分利用过程数据所表征的工业运行过程的特性<sup>[163]</sup>. 因此, 借鉴半监督学习思想, 在虚拟样本

输入生成阶段可充分利用未标记过程数据所能表征的特征空间以提高生成样本的质量. 结合上述样本的差异度与主动学习算法筛选合格输入数据, 对其进行标记能够获得高置信度的伪标记样本和高质量的虚拟样本. 笔者认为, 从输入输出视角, 真实样本可记为“真-真”样本, 之前研究所生成的虚拟样本可记为“虚-虚”样本, 此处采用半监督方式获得的样本可记为“真-虚”样本. 因此, 基于监督和半监督学习的集成 VSG 能够基于“真-真”样本和未标记样本提高“虚-虚”样本可信度的同时通过“真-虚”样本进一步增加虚拟样本的数量.

#### 7) 自适应更新的动态 VSG

在实际的工业过程中, 数据分布会随时间发生动态变化导致旧模型无法适用于新样本, 该问题被称为概念漂移, 产生原因通常是元器件老化或生产环境变化导致模型输入输出间的分布关系发生改变. 如何进行概念漂移的检测、量化和处理也是学术界的开放性课题之一<sup>[164]</sup>. 基于历史真实数据的 VSG 虽能够进行域扩展, 但却难以表征工业过程未知漂移和难以确定未知域. 因此, VSG 应能够根据工业动态环境的变化进行完备样本分布的实时更新, 进而确保生成模型的性能和预测模型的精度, 在该方向上的研究成果还未见报道.

## 5 结论

本文总结了针对复杂工业过程难测运行指标和异常故障进行建模的真实样本所存在的问题, 梳理了虚拟样本的定义和内涵, 给出了工业过程 VSG 的实现流程, 综述了面向样本覆盖区域、实现流程与推广应用 3 个方向的研究现状, 讨论了未来研究方向. 结合上述分析结果, 笔者认为未来挑战包括: 1) 构建合成数据集进行 VSG 理论分析, 进行样本质量与生成模型的协同优化; 2) 利用对抗学习对机理知识、经验规则和数据驱动模型进行动态进化选择, 构建具有最优生成流程的智能 VSG; 3) 同时从输入和输出角度评估本文所提出的相似域样本空间, 采用基于样本和模型的迁移学习构建虚拟样本输入生成模型和输出映射模型; 4) 面向工业过程的物理实体构建混合机理和数据驱动的数字孪生系统, 依据实际工业数据的动态变化对数据孪生模型进行预测性调整以确保虚拟样本质量和预测模型性能; 5) 利用未标记样本提升虚拟样本的可信度, 结合监督和半监督学习算法的差异度和主动学习算法的灵活性, 构建面向多视角学习机制的集成 VSG 和结合工业过程概念漂移的动态 VSG.

## 附录 A

表 A1 VSG 的研究成果统计与对比  
Table A1 Statistics and comparison of VSG research results

分类	子分类	方法	年份	优劣	文献
面向样本覆盖区域之原始域样本空间回归 VSG	特征工程	LLE + BPNN	2020		[52]
		Isomap + 插值法	2020	特征变换, 流形学习更加直观, 但特征失去物理含义	[53]
		<i>t</i> -SNE + RF	2021		[54]
		机理	2021	特征选择, 工业过程知识获取困难	[55]
		两者结合	2020	综合特征变换与选择, 具有较强的定制化特性	[56]
	样本工程	空间投影 + RBF	2021	函数模型, 空间投影具有新颖性	[57]
		数据趋势	2021	函数模型, 提出的稀疏假设和集中假设具有参考价值	[49]
		总线拓扑结构插值	2023	函数模型, 有效控制插值位置	[58]
		RWNN 插值法	2018	函数模型, 基于神经网络	[59]
		AANN 插值	2019	模型学习样本的非线性分布关系	[60]
		RWNN + 等间隔插值法	2020	对小样本难以有效	[56]
		MTD + PSO	2021	函数模型, PSO 优化选择虚拟样本	[61]
		多目标 PSO	2022	函数模型, 多目标 PSO 优化选择虚拟样本和生成数量	[15]
		LOF + <i>K</i> -means + GAN	2021	对抗模型, 插值生成输出, CGAN 生成输入	[20]
		双 GAN	2022	对抗模型, 两种 GAN 分别负责输入和输出的生成, 复杂性高	[64]
回归器 + CWGAN	2022	对抗模型, 通过回归器匹配虚拟样本输出并共同训练	[65]		
面向样本覆盖区域之原始域样本空间的分类 VSG	特征工程	添加编码器	2020	采用编码器提取特征	[68]
		添加卷积层	2021	添加卷积层提取特征	[66]
		添加卷积层	2021		[67]
		添加自注意	2022	添加自注意力模型增强特征	[69]

表 A1 VSG 的研究成果统计与对比 (续表)

Table A1 Statistics and comparison of VSG research results (continued table)

分类	子分类	方法	年份	优劣	文献
面向样本覆盖区域之原始域样本空间的分类 VSG	样本工程	基于加权核的 SMOTE	2018	函数模型, 解决 SMOTE 算法在高 IR 下的非线性可分离问题	[71]
		Minkowski 距离替换欧氏距离	2019	函数模型, 有效生成高维虚拟样本	[72]
		SMOTE + 决策树	2020	函数模型, 决策树算法提取关键规则	[73]
		SMOTE + SVM	2020	函数模型, 支持向量边界生成虚拟样本	[74]
		范围控制 SMOTE	2021	函数模型, 有效地缓解范围偏移和边界样本重叠等问题	[75]
		超球面空间 + 组发现技术	2007	函数模型, 由数据的结构生成虚拟样本	[76]
		组发现技术 + 纯化过程	2020	函数模型, 纯化过程剔除冗余样本	[77]
		ACGAN	2020	对抗模型, 添加 Dropout 层防止过拟合, 添加卷积层提取更多特征	[79]
		ACWGAN-GP	2020	对抗模型, ACGAN 的进化版	[80]
		MAML + ACGAN	2021	对抗模型, MAML 初始化和更新网络使得生成过程更加稳定	[81]
		GAN + 多尺度 CNN	2021	对抗模型, 生成模型需要改进	[82]
		DCGAN + K-means	2021	对抗模型, K-means 算法对模型改进	[67]
		MoGAN	2021	对抗模型, 判别器既判断样本真假又充当分类器和故障检测器	[83]
		GAN + MSCNN	2021	对抗模型, 多 GAN 联合生成	[82]
		贝叶斯优化 + WGAN	2021	对抗模型, 贝叶斯优化策略自适应调节判别器参数	[86]
		WGAN + LSTM-FCN	2022	对抗模型, 结合 LSTM	[84]
		AE + GAN	2019	对抗模型, AE 结合 GAN	[89]
		VAE + GAN	2020		[68]
		深度残差网络 + VAE + GAN	2021	对抗模型, 深度残差网络提高模型性能	[90]
		AE + LSGAN	2022	对抗模型, 暹罗编码器计算特征残差	[91]
CVAEGAN-SM	2022	对抗模型, 生成器中加入自调制机制	[92]		
堆叠 AE + WGAN	2023	对抗模型, 提升了模型的生成能力	[93]		
面向样本覆盖区域之扩展域样本空间的回归 VSG	集合理论	正态隶属度	1997	模糊集理论, 仅适用于扩展范围对称情况	[94]
		DNN	2003	模糊集理论, 特征相关系数大于 0.9 才能计算扩展范围	[95]
		MTD	2007	模糊集理论, 通过假设特征独立不对称地扩散特征范围	[96]
		GTD	2010	模糊集理论, 增量版的 MTD	[97]
		TTD	2012	模糊集理论, 与树算法结合	[98]
		神经网络 MTD	2012	模糊集理论, 神经网络与 MTD 结合	[99]
		MD-MTD	2016	模糊集理论, 三角和均匀分布组合的多分布	[100]
		KNN + MTD	2022	模糊集理论, KNN 确保合理的扩展范围	[101]
		K-means + MTD	2022	模糊集理论, K-means 解决属性冗余	[102]
		AD-MTD + MD-MTD	2019	模糊集理论, 多种算法结合取长补短	[103]
		MTD + RWNN	2020	模糊集理论, 改进分布 + 隐含层插值	[56]
		MTD + GA	2014	模糊集理论, 基于优化算法搜寻虚拟样本, 更合理	[46]
		TMIE + PSO	2016	模糊集理论, PSO 优化选择虚拟样本	[104]
		MTD + PSO	2021		[61]
		分布假设		IKDE	2006
时序 IKDE	2008			高斯分布, 用于时序数据	[110]
SJDT	2016			高斯分布, SJDT 将数据趋于正态分布	[111]
MPV	2013			非高斯分布, 多样本分布	[112]
假设检验	2019			非高斯分布, 先聚类再估计	[113]
基于知识		多目标 PSO	2022	基于知识确定输出扩展域下限, 多目标 PSO 优化选择虚拟样本和生成数量	[15]

表 A1 VSG 的研究成果统计与对比 (续表)

Table A1 Statistics and comparison of VSG research results (continued table)

分类	子分类	方法	年份	优劣	文献	
面向样本覆盖区域之 扩展域样本空间的 分类 VSG	集合理论	FID	2017	模糊集理论, 既生成虚拟样本又填充缺失	[114]	
		SMOTE + 粗糙集理论	2012	粗糙集理论, 扩展范围有限	[115]	
		三支决策	2018	粗糙集理论, 未精准计算扩展范围	[116]	
	分布假设	假设分布	2010	高斯分布, 计算数据的均值和方差确定高斯分布	[43]	
		假设分布	2022	高斯分布, AIC 和 BIC 自适应确定高斯分布参数	[117]	
		SVM	2013	非高斯分布, 状态函数采样生成虚拟样本	[118]	
		K-means + Weibull 分布	2014	非高斯分布, 特定过程采用特定分布	[119]	
	基于知识	FAGAN	2021	基于知识, 专家知识定义的故障属性作为辅助信息以 使得生成样本	[123]	
	面向 VSG 实现流程之 回归问题	过程数据预处理阶段	缺失值删减和人工填充	2021	有效减少缺失和异常值对数据的影响但会减少样本数量	[61]
				2022		[15]
			2020	[127]		
缺失和异常值识别剔除			2022	[64]		
LLE			2020	流形学习更加直观, 特征失去物理含义	[52]	
Isomap		2020	[53]			
t-SNE		2021	[54]			
根据化工机理选择特征		2017	机理知识获取困难	[128]		
		2018		[59]		
专家经验		2021	特定实验知识	[61]		
虚拟样本输入 生成阶段		欧氏距离识别稀疏区域	2020	引入欧氏距离	[127]	
		投影最大间距识别稀疏	2021	引入投影最大间距	[57]	
		可视化样本分布识别稀疏 区域		2020	可视化, 直观	[52]
				2020		[53]
				2021		[54]
	稀疏性和集中性假设	2021	确定稀疏和密集区域关系	[49]		
	WGAN-GP	2022	引入 GAN 用于回归	[64]		
	CWGAN	2022		[65]		
	MTD	2007	确定虚拟样本输入的扩展域范围后插值	[96]		
	TMIE + PSO	2016		[104]		
流形子空间 + MTD	2021	[129]				
虚拟样本输出 生成阶段		2018	映射模型的性能受限于小样本	[59]		
	RWNN 映射模型	2016		[104]		
		2021		[129]		
	BPNN 映射模型	2020		[52]		
	RF 映射模型	2021	[54]			
	RBF 映射模型	2021	[57]			
	CS-CGAN 匹配输出	2022	匹配模型与虚拟样本输入同时训练	[64]		
	回归器匹配输出	2022		[65]		
分位数回归器匹配输出	2021	[55]				
虚拟样本质量 筛选阶段	模型误差小于 10% 筛选	2014	受限于小样本建模性能	[46]		
	隶属度的似然估计筛选	2018	引入似然估计	[130]		
	PSO 优化算法筛选	2021	引入优化算法	[61]		
	专家筛选	2021	具有主观性	[49]		
虚拟样本数量 确定阶段	信息熵	2019	引入信息熵	[131]		
	稀疏和集中假设	2021	引入各种假设	[49]		

表 A1 VSG 的研究成果统计与对比 (续表)

Table A1 Statistics and comparison of VSG research results (continued table)

分类	子分类	方法	年份	优劣	文献	
面向 VSG 实现流程之 分类问题	特殊阶段	LOF + CGAN	2021	先生成虚拟样本输出后匹配虚拟样本输入	[20]	
		三样条插值 + ITNN	2021		[49]	
	过程数据预处理阶段	信号数据转换为灰度图		2022	借鉴图像领域算法处理	[85]
				2022		[132]
		重叠分割、旋转和抖动的数据增强		2021	缓解过拟合	[133]
		SMOTE + 粗糙集理论		2012	扩展范围有限	[115]
		SMOTE + 决策树		2020	决策树算法提取运行规则	[73]
		SMOTE + SVM		2020	引入支持向量机边界	[74]
		Minkowski 距离替换欧氏距离		2019	可以有效地生成高维虚拟样本	[72]
		范围控制 SMOTE		2021	通过控制生成范围减少边界重叠样本	[75]
		WGAN + LSTM-FCN		2022	引入 LSTM	[84]
		CWGAN-GP + FDGRU		2022	增加梯度惩罚项和条件信息	[85]
		Pull-away 损失函数 GAN		2022	添加自注意力模型增强特征	[69]
	虚拟样本输入生成阶段	ACGAN		2020	添加 Dropout 层防止过拟合, 添加卷积层提取更多特征	[79]
				2020		[68]
		深度残差网络 + VAE + GAN		2021	深度残差网络提高模型性能	[90]
		MoGAN		2021	判别器包含真假判断、故障诊断和故障分类三种功能	[83]
		CVAEGAN-SM		2022	生成器加入自调制机制	[92]
		并行 GAN		2020	对应多类别同时训练, 复杂性高	[19]
		SMOTE + VAE		2018		[134]
		自适应混合		2020	基于样本的迁移学习 VSG	[135]
		迁移学习 + 插值		2022		[136]
		Fine-tuning + WGAN		2021		[137]
		迁移学习 + GAN		2022	基于模型的迁移学习 VSG	[138]
			Wasserstein 距离	2020		[139]
			KL 散度, F-score, Kappa 系数, GAN 测试值	2021	未给出评价指标的具体限值筛选虚拟样本	[66]
		虚拟样本质量筛选阶段	Wasserstein 距离, KL 散度, 欧氏距离, 皮尔逊相关系数	2021		[67]
	马氏距离, 欧氏距离		2021		[82]	
		判别概率, 最大均值差异, KL 散度	2022	给出评价指标的具体限值筛选虚拟样本	[69]	
		皮尔逊相关系数	2022		[85]	
		最大均值差异, KL 散度, GAN 测试值	2022	未给出评价指标的具体限值筛选虚拟样本	[92]	
	虚拟样本数量确定阶段	分类复杂度确定虚拟样本数量	1998	采用分类复杂度确定虚拟样本数量	[14]	
面向 VSG 推广应用的 回归问题		TMIE + PSO	2016	PSO 优化选择	[104]	
		RWNN 插值法	2018		[59]	
	石油化工	Isomap + 插值法		2020	提出隐含层插值生成虚拟样本	[53]
		分位数回归器匹配输出		2021	提出分位数回归匹配输出	[55]
		回归器 + CWGAN		2022	通过回归器匹配虚拟样本输出并同时训练	[65]
		两者结合		2020	具有较强的定制化特性	[56]
	固废焚烧	MTD + PSO		2021	PSO 优化选择虚拟样本	[61]
多目标 PSO			2022	多目标 PSO 优化选择虚拟样本和生成数量	[15]	

表 A1 VSG 的研究成果统计与对比 (续表)

Table A1 Statistics and comparison of VSG research results (continued table)

分类	子分类	方法	年份	优劣	文献
面向 VSG 推广应用的分类问题	工业制造	GTD	2010	增量版的 MTD	[97]
		TTD	2012	与树算法结合	[98]
		MPV	2013	采用多分布	[112]
		模糊 $c$ 均值聚类 + 箱线图	2018	箱线图确定扩展范围	[141]
		假设分布	2022	AIC 和 BIC 自适应确定高斯分布参数	[117]
	矿业冶金	时频变换 + FBP + 信息熵	2018	特定问题采用特定方法	[6]
		RWNN 插值 + MD-MTD	2019	GA 优化选择虚拟混合样本	[145]
	滚动轴承故障诊断	迁移学习 + GAN	2020	迁移与 GAN 相结合	[146]
		PGDAE + DCN	2021	引入 PGDAE	[147]
		元学习 + WAE	2021	元学习提高虚拟样本质量	[66]
		CVAEGAN-SM	2022	生成器加入自调制机制	[92]
		DSAN	2022	自注意模块增强深度特征	[132]
		GAN	2022	常数 $Q$ 转换将信号转换为频谱图, 均方差替换交叉熵	[148]
		ACGAN	2022	引入 ACGAN	[149]
		特征增强 GAN	2022	自注意模块增强深度特征	[69]
		DFGN	2021	可用于零样本故障诊断	[150]
	变压器故障诊断	SMOTE + 决策树	2020	决策树算法提取关键规则	[73]
		SMOTE + SVM	2020	提出支持向量边界生成样本	[74]
		CWGAN-GP	2020	引入梯度惩罚	[151]
		AE + LSGAN	2022	暹罗编码器计算特征残差	[91]
涡轮机故障诊断	GAN	2019		[152]	
	DACNN	2019	结合 GAN 与具体问题	[153]	
	VAE + GAN	2019		[89]	
	ID-CNN GAN	2019	虚拟样本输出和故障诊断组合模型	[154]	
齿轮箱故障诊断	ACGAN + CVAE	2020	引入 CVAE	[68]	
	贝叶斯优化 + WGAN	2021	贝叶斯优化策略自适应调节判别器参数	[86]	
	DCGAN + $K$ -means	2021	$K$ -means 算法对模型改进	[67]	

表 A2 符号说明

Table A2 Symbol description

缩写词	英文全称	中文全称
VSG	Virtual sample generation	虚拟样本生成
MSWI	Municipal solid waste incineration	城市固废焚烧
DXN	Dioxin	二噁英
VAE	Variational autoencoder	变分自编码器
GAN	Generative adversarial network	生成对抗网络
FDD	Fault detection and diagnosis	故障检测与诊断
IR	Imbalance ratio	不平衡比
SMOTE	Synthetic minority over-sampling technique	合成少数类过采样技术
MAPE	Mean absolute percentage error	平均绝对百分比误差
LLE	Locally linear embedding	局部线性嵌入
BPNN	Back propagation neural network	反向传播神经网络
Isomap	Isometric feature mapping	等距特征映射
$t$ -SNE	$t$ -distributed stochastic neighbor embedding	$t$ 分布随机邻域嵌入

表 A2 符号说明 (续表)

Table A2 Symbol description (continued table)

缩写词	英文全称	中文全称
RF	Random forest	随机森林
RBF	Radial basis function	径向基函数
CSI	Cubic spline interpolation	三样条插值
ITNN	Input-training neural network	输入训练神经网络
RWNN	Random weight neural network	随机权神经网络
AANN	Auto-associative neural network	自联想神经网络
LOF	Local outlier factor	局部异常因子
CGAN	Conditional generative adversarial network	条件生成对抗网络
CS-CGAN	Cycle structure conditional generative adversarial network	循环结构条件生成对抗网络
FFT	Fast Fourier transform	快速傅里叶变换
SVM	Support vector machine	支持向量机
AC-GAN	Auxiliary classifier generative adversarial network	辅助分类器生成对抗网络
ACWGAN-GP	Auxiliary classifier Wasserstein generative adversarial network with gradient penalty	具有梯度惩罚的辅助分类 Wasserstein 生成对抗网络
MAML	Model agnostic meta learning	模型无关元学习
CNN	Convolutional neural network	卷积神经网络
DCGAN	Deep convolutional generative adversarial network	深度卷积生成对抗网络
MoGAN	Minority oversampling generative adversarial network	少数类过采样生成对抗网络
AE	Autoencoder	自编码器
CVAE-GAN	Conditional variational autoencoder generative adversarial network	条件变分自编码器生成对抗网络
LSGAN	Least squares generative adversarial network	最小二乘生成对抗网络
DNN	Diffusion neural network	扩散神经网络
MTD	Mega-trend-diffusion	大趋势扩散
GTD	Generalized-trend-diffusion	广义趋势扩散
TTD	Tree structure based trend diffusion	树结构趋势扩散
MD-MTD	Multi-distribution mega-trend-diffusion	多分布大趋势扩散
KNN	$K$ -nearest neighbor	$K$ 近邻
AD-MTD	Advanced mega-trend-diffusion	改进型大趋势扩散
Hybrid-MTD	Hybrid mega-trend-diffusion	混合大趋势扩散
GA	Genetic algorithm	遗传算法
FBP	Feasibility-based programming	可行性的规划
TMIE	Information-expanded based on triangular membership	基于三角隶属度的信息扩散
PSO	Particle swarm optimization	粒子群优化
IKDE	Improved kernel density estimation	改善核密度估计
SJDT	Small Johnson data transformation	小型约翰变换方法
MPV	Maximal $p$ value	最大 $p$ 值
AIC	Akaike information criterion	赤池信息准则
AICc	Corrected version of the akaike information criterion	修正版赤池信息准则
FID	Fuzzy-based information decomposition	基于模糊的信息分解
BIC	Bayesian information criterion	贝叶斯信息准则
FAGAN	Fault attributes generative adversarial network	故障属性生成对抗网络

表 A2 符号说明 (续表)

Table A2 Symbol description (continued table)

缩写词	英文全称	中文全称
SRWGAN	Semantic refinement Wasserstein generative adversarial network	语义细化 Wasserstein 生成对抗网络
MOPSO	Multi-objective particle swarm optimization	多目标粒子群优化
PGDAE	Predictive generative denoising autoencoder	预测生成去噪自编码器
DCN	Deep coral network	深度珊瑚网络
WAE	Wasserstein autoencoder	Wasserstein 自编码器
DSAN	Deep subdomain adaptation network	深度子域适应网络
DFGN	Deep feature generating network	深度特征生成网络
DACNN	Deep adversarial convolutional neural network	深度对抗卷积神经网络
BO	Bayesian optimization	贝叶斯优化
DCGAN	Deep convolution generative adversarial network	深度卷积生成对抗网络

## References

- Yin S, Ding S X, Xie X C, Luo H. A review on basic data-driven approaches for industrial process monitoring. *IEEE Transactions on Industrial Electronics*, 2014, **61**(11): 6418–6428
- Sun Y N, Zhuang Z L, Xu H W, Qin W, Feng M J. Data-driven modeling and analysis based on complex network for multimode recognition of industrial processes. *Journal of Manufacturing Systems*, 2022, **62**: 915–924
- Chai Tian-You. Industrial process control systems: Research status and development direction. *Scientia Sinica Informationis*, 2016, **46**(8): 1003–1015  
(柴天佑. 工业过程控制系统研究现状与发展方向. 中国科学: 信息科学, 2016, **46**(8): 1003–1015)
- Ding Jin-Liang, Yang Cui-E, Chen Yuan-Dong, Chai Tian-You. Research progress and prospects of intelligent optimization decision making in complex industrial process. *Acta Automatica Sinica*, 2018, **44**(11): 1931–1943  
(丁进良, 杨翠娥, 陈远东, 柴天佑. 复杂工业过程智能优化决策系统的现状与展望. 自动化学报, 2018, **44**(11): 1931–1943)
- Xia H, Tang J, Qiao J F, Zhang J, Yu W. DF classification algorithm for constructing a small sample size of data-oriented DF regression model. *Neural Computing and Applications*, 2022, **34**(4): 2785–2810
- Tang Jian, Qiao Jun-Fei, Chai Tian-You, Liu Zhuo, Wu Zhi-Wei. Modeling multiple components mechanical signals by means of virtual sample generation technique. *Acta Automatica Sinica*, 2018, **44**(9): 1569–1589  
(汤健, 乔俊飞, 柴天佑, 刘卓, 吴志伟. 基于虚拟样本生成技术的多组分机械信号建模. 自动化学报, 2018, **44**(9): 1569–1589)
- Ma Da-Zhong, Hu Xu-Guang, Sun Qiu-Ye, Zheng Jun, Wang Rui. Cyber-physical abnormality diagnosis method using data feature fusion for pipeline network. *Acta Automatica Sinica*, 2019, **45**(1): 163–173  
(马大中, 胡旭光, 孙秋野, 郑君, 王睿. 基于数据特征融合的管网信息物理异常诊断方法. 自动化学报, 2019, **45**(1): 163–173)
- Qiao Jun-Fei, Guo Zi-Hao, Tang Jian. Dioxin emission concentration measurement approaches for municipal solid wastes incineration process: A survey. *Acta Automatica Sinica*, 2020, **46**(6): 1063–1089  
(乔俊飞, 郭子豪, 汤健. 面向城市固废焚烧过程的二噁英排放浓度检测方法综述. 自动化学报, 2020, **46**(6): 1063–1089)
- Xia H, Tang J, Aljerf L. Dioxin emission prediction based on improved deep forest regression for municipal solid waste incineration process. *Chemosphere*, 2022, **294**: Article No. 133716
- Poggio T, Vetter T. Recognition and Structure from One 2D Model View: Observations on Prototypes, Object Classes and Symmetries. C.B.L.P. Paper No. 69, Massachusetts Institute of Technology, United States, 1992.
- Liu X F, Sun Q Q, Meng Y, Fu M, Bourenmane S. Hyperspectral image classification based on parameter-optimized 3D-CNNs combined with transfer learning and virtual samples. *Remote Sensing*, 2018, **10**(9): Article No. 1425
- Li L J, Peng Y L, Qiu G Y, Sun Z G, Liu S G. A survey of virtual sample generation technology for face recognition. *Artificial Intelligence Review*, 2018, **50**(1): 1–20
- Zhao X C, Lv W M. Reliability evaluation of complex equipment based on virtual samples and performance degradation. In: Proceedings of the International Conference on Electronics and Electrical Engineering Technology. Tianjin, China: ACM, 2018. 112–117
- Niyogi P, Girosi F, Poggio T. Incorporating prior information in machine learning by creating virtual examples. *Proceedings of the IEEE*, 1998, **86**(11): 2196–2209
- Wang Dan-Dan, Tang Jian, Xia Heng, Qiao Jun-Fei. Virtual sample generation method based on hybrid optimization with multi-objective PSO. *Acta Automatica Sinica*, DOI: 10.16383/j.aas.c211091  
(王丹丹, 汤健, 夏恒, 乔俊飞. 基于多目标 PSO 混合优化的虚拟样本生成. 自动化学报, DOI: 10.16383/j.aas.c211091)
- Kingma D P, Welling M. Auto-encoding variational Bayes. In: Proceedings of the 2nd International Conference on Learning Representations. Banff, Canada: ICLR, 2014.
- Zhang L, Zhang H, Cai G W. The multiclass fault diagnosis of wind turbine bearing based on multisource signal fusion and deep learning generative model. *IEEE Transactions on Instrumentation and Measurement*, 2022, **71**: Article No. 3514212
- Wang X, Liu H. Data supplement for a soft sensor using a new generative model based on a variational autoencoder and wasserstein GAN. *Journal of Process Control*, 2020, **85**: 91–99
- Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, et al. Generative adversarial networks. *Communications of the ACM*, 2020, **63**(11): 139–144
- Zhu Q X, Hou K R, Chen Z S, Gao Z S, Xu Y, He Y L. Novel virtual sample generation using conditional GAN for developing soft sensor with small data. *Engineering Applications of Artificial Intelligence*, 2021, **106**: Article No. 104497
- Li Yan-Rui, Yang Chun-Jie, Zhang Han-Wen, Li Jun-Fang. Discussion on key technologies of digital twin in process industry. *Acta Automatica Sinica*, 2021, **47**(3): 501–514  
(李彦瑞, 杨春节, 张瀚文, 李俊方. 流程工业数字孪生关键技术探讨. 自动化学报, 2021, **47**(3): 501–514)
- Yang Lin-Yao, Chen Si-Yuan, Wang Xiao, Zhang Jun, Wang Cheng-Hong. Digital twins and parallel systems: State of the art, comparisons and prospect. *Acta Automatica Sinica*, 2019, **45**(11): 2001–2031  
(杨林瑶, 陈思源, 王晓, 张俊, 王成红. 数字孪生与并行系统: 发

- 展现状、对比及展望. 自动化学报, 2019, **45**(11): 2001–2031)
- 23 Lee L H, Braud T, Zhou P Y, Wang L, Xu D L, Lin Z J, et al. All one needs to know about metaverse: A complete survey on technological singularity, virtual ecosystem, and research agenda. arXiv: 2110.05352, 2021.
- 24 Ning H S, Wang H, Lin Y J, Wang W X, Dhelim S, Farha F, et al. A survey on metaverse: The state-of-the-art, technologies, applications, and challenges. arXiv: 2111.09673, 2021.
- 25 Yao L, Ge Z Q. Industrial big data modeling and monitoring framework for plant-wide processes. *IEEE Transactions on Industrial Informatics*, 2021, **17**(9): 6399–6408
- 26 Ma X, Si Y B, Qin Y H, Wang Y Q. Fault detection for dynamic processes based on recursive innovational component statistical analysis. *IEEE Transactions on Automation Science and Engineering*, 2023, **20**(1): 310–319
- 27 Ma X, Wu D H, Gao S X, Hou T Z, Wang Y Q. Autocorrelation feature analysis for dynamic process monitoring of thermal power plants. *IEEE Transactions on Cybernetics*, 2023, **53**(8): 5387–5399
- 28 Xia Heng, Tang Jian, Cui Can-Lin, Qiao Jun-Fei. Soft sensing method of dioxin emission in municipal solid waste incineration process based on broad hybrid forest regression. *Acta Automatica Sinica*, 2023, **49**(2): 343–365  
(夏恒, 汤健, 崔璨麟, 乔俊飞. 基于宽度混合森林回归的城市固废焚烧过程二噁英排放浓度软测量. 自动化学报, 2023, **49**(2): 343–365)
- 29 Cui M L, Wang Y Q, Lin X S, Zhong M Y. Fault diagnosis of rolling bearings based on an improved stack autoencoder and support vector machine. *IEEE Sensors Journal*, 2021, **21**(4): 4927–4937
- 30 Md Nor N, Che Hassan C R, Hussain M A. A review of data-driven fault detection and diagnosis methods: Applications in chemical process systems. *Reviews in Chemical Engineering*, 2020, **36**(4): 513–553
- 31 Raudys Š. Trainable fusion rules. II. Small sample-size effects. *Neural Networks*, 2006, **19**(10): 1517–1527
- 32 Tang J, Jia M Y, Liu Z, Chai T Y, Yu W. Modeling high dimensional frequency spectral data based on virtual sample generation technique. In: Proceedings of the IEEE International Conference on Information and Automation. Lijiang, China: IEEE, 2015. 1090–1095
- 33 Zhong K, Han M, Han B. Data-driven based fault prognosis for industrial systems: A concise overview. *IEEE/CAA Journal of Automatica Sinica*, 2019, **7**(2): 330–345
- 34 Ji C, Sun W. A review on data-driven process monitoring methods: Characterization and mining of industrial data. *Processes*, 2022, **10**(2): Article No. 335
- 35 Zhu R, Guo Y W, Xue J H. Adjusting the imbalance ratio by the dimensionality of imbalanced data. *Pattern Recognition Letters*, 2020, **133**: 217–223
- 36 Martin-Diaz I, Morinigo-Sotelo D, Duque-Perez O, de J. Romero-Troncoso R. Early fault detection in induction motors using AdaBoost with imbalanced small data and optimized sampling. *IEEE Transactions on Industry Applications*, 2017, **53**(3): 3066–3075
- 37 Feng L J, Zhao C H. Fault description based attribute transfer for zero-sample industrial fault diagnosis. *IEEE Transactions on Industrial Informatics*, 2021, **17**(3): 1852–1862
- 38 Yu W K, Zhao C H. Broad convolutional neural network based industrial process fault diagnosis with incremental learning capability. *IEEE Transactions on Industrial Electronics*, 2020, **67**(6): 5081–5091
- 39 Lou Z J, Wang Y Q, Si Y B, Lu S. A novel multivariate statistical process monitoring algorithm: Orthonormal subspace analysis. *Automatica*, 2022, **138**: Article No. 110148
- 40 Wang K, Li J, Tsung F. Distribution inference from early-stage stationary data streams by transfer learning. *IIESE Transactions*, 2022, **54**(3): 303–320
- 41 Zhou X F, Zhai N J, Li S, Shi H B. Time series prediction method of industrial process with limited data based on transfer learning. *IEEE Transactions on Industrial Informatics*, 2023, **19**(5): 6872–6882
- 42 Maschler B, Weyrich M. Deep transfer learning for industrial automation: A review and discussion of new techniques for data-driven machine learning. *IEEE Industrial Electronics Magazine*, 2021, **15**(2): 65–75
- 43 Yang J, Yu X, Xie Z Q, Zhang J P. A novel virtual sample generation method based on Gaussian distribution. *Knowledge-Based Systems*, 2011, **24**(6): 740–748
- 44 Guo Zi-Hao. Soft Sensing of Dioxin Emission Concentration for Municipal Solid Waste Incineration [Master thesis], Beijing University of Technology, China, 2021.  
(郭子豪. 面向城市固废焚烧过程的二噁英排放浓度软测量 [硕士学位论文], 北京工业大学, 中国, 2021.)
- 45 Tang J, Xia H, Aljerf L, Wang D D, Ukaogo P O. Prediction of dioxin emission from municipal solid waste incineration based on expansion, interpolation, and selection for small samples. *Journal of Environmental Chemical Engineering*, 2022, **10**(5): Article No. 108314
- 46 Li D C, Wen I H. A genetic algorithm-based virtual sample generation technique to improve small data set learning. *Neurocomputing*, 2014, **143**: 222–230
- 47 Ren Y F, Liu J H, Zhang H G, Wang J F. TBDA-Net: A task-based bias domain adaptation network under industrial small samples. *IEEE Transactions on Industrial Informatics*, 2022, **18**(9): 6109–6119
- 48 Fu M R, Liu J H, Zhang H G, Lu S X. Multisensor fusion for magnetic flux leakage defect characterization under information incompleteness. *IEEE Transactions on Industrial Electronics*, 2021, **68**(5): 4382–4392
- 49 Chen Z S, Zhu Q X, Xu Y, He Y L, Su Q L, Liu Y C, et al. Integrating virtual sample generation with input-training neural network for solving small sample size problems: Application to purified terephthalic acid solvent system. *Soft Computing*, 2021, **25**(8): 6489–6504
- 50 Zhao Chun-Hui, Hu Yun-Yun, Zheng Jia-Le, Chen Jun-Hao. Data-driven operating monitoring for coal-fired power generation equipment: The state of the art and challenge. *Acta Automatica Sinica*, 2022, **48**(11): 2611–2633  
(赵春晖, 胡赞响, 郑嘉乐, 陈军豪. 数据驱动的燃煤发电装备运行工况监控——现状与展望. 自动化学报, 2022, **48**(11): 2611–2633)
- 51 Zhu J L, Ge Z Q, Song Z H, Gao F R. Review and big data perspectives on robust data mining approaches for industrial process modeling with outliers and missing data. *Annual Reviews in Control*, 2018, **46**: 107–133
- 52 Zhu Q X, Zhang X H, He Y L. Novel virtual sample generation based on locally linear embedding for optimizing the small sample problem: Case of soft sensor applications. *Industrial & Engineering Chemistry Research*, 2020, **59**(40): 17977–17986
- 53 Zhang X H, Xu Y, He Y L, Zhu Q X. Novel manifold learning based virtual sample generation for optimizing soft sensor with small data. *ISA Transactions*, 2021, **109**: 229–241
- 54 He Y L, Hua Q, Zhu Q X, Lu S. Enhanced virtual sample generation based on manifold features: Applications to developing soft sensor using small data. *ISA Transactions*, 2022, **126**: 398–406
- 55 Chen Zhong-Sheng, Zhu Mei-Yu, He Yan-Lin, Xu Yuan, Zhu Qun-Xiong. Quantile regression CGAN based virtual samples generation and its applications to process modeling. *CIESC Journal*, 2021, **72**(3): 1529–1538  
(陈忠圣, 朱梅玉, 贺彦林, 徐圆, 朱群雄. 基于分位数回归 CGAN 的虚拟样本生成方法及其过程建模应用. 化工学报, 2021, **72**(3): 1529–1538)
- 56 Qiao Jun-Fei, Guo Zi-Hao, Tang Jian. Virtual sample generation method based on improved megatrend diffusion and hidden layer interpolation with its application. *CIESC Journal*, 2020, **71**(12): 5681–5695

- (乔俊飞, 郭子豪, 汤健. 基于改进大趋势扩散和隐含层插值的虚拟样本生成方法及应用. *化工学报*, 2020, **71**(12): 5681–5695)
- 57 Zhu Q X, Liu D P, Xu Y, He Y L. Novel space projection interpolation based virtual sample generation for solving the small data problem in developing soft sensor. *Chemometrics and Intelligent Laboratory Systems*, 2021, **217**: Article No. 104425
- 58 Sutojo T, Rustad S, Akrom M, Syukur A, Shidik G F, Dipojono H K. A machine learning approach for corrosion small datasets. *Npj Materials Degradation*, 2023, **7**(1): Article No. 18
- 59 He Y L, Wang P J, Zhang M Q, Zhu Q X, Xu Y. A novel and effective nonlinear interpolation virtual sample generation method for enhancing energy prediction and analysis on small data problem: A case study of Ethylene industry. *Energy*, 2018, **147**: 418–427
- 60 Zhu Bao, Qiao Jun-Fei. Novel virtual sample generation based on feature scaling of auto-associative neural network and its applications to process modeling. *Computers and Applied Chemistry*, 2019, **36**(4): 304–307  
(朱宝, 乔俊飞. 基于 AANN 特征缩放的虚拟样本生成方法及其过程建模应用. *计算机与应用化学*, 2019, **36**(4): 304–307)
- 61 Tang Jian, Wang Dan-Dan, Guo Zi-Hao, Qiao Jun-Fei. Prediction of dioxin emission concentration in the municipal solid waste incineration process based on optimal selection of virtual samples. *Journal of Beijing University of Technology*, 2021, **47**(5): 431–443  
(汤健, 王丹丹, 郭子豪, 乔俊飞. 基于虚拟样本优化选择的城市固废焚烧过程二噁英排放浓度预测. *北京工业大学学报*, 2021, **47**(5): 431–443)
- 62 Ren Hao, Qu Jian-Feng, Chai Yi, Tang Qiu, Ye Xin. Deep learning for fault diagnosis: The state of the art and challenge. *Control and Decision*, 2017, **32**(8): 1345–1358  
(任浩, 屈剑锋, 柴毅, 唐秋, 叶欣. 深度学习在故障诊断领域中的研究现状与挑战. *控制与决策*, 2017, **32**(8): 1345–1358)
- 63 Hu Ming-Fei, Zuo Xin, Liu Jian-Wei. Survey on deep generative model. *Acta Automatica Sinica*, 2022, **48**(1): 40–74  
(胡铭菲, 左信, 刘建伟. 深度生成模型综述. *自动化学报*, 2022, **48**(1): 40–74)
- 64 Zhu Q X, Xu T X, Xu Y, He Y L. Improved virtual sample generation method using enhanced conditional generative adversarial networks with cycle structures for soft sensors with limited data. *Industrial & Engineering Chemistry Research*, 2022, **61**(1): 530–540
- 65 He Y L, Li X Y, Ma J Y, Lu S, Zhu Q X. A novel virtual sample generation method based on a modified conditional wasserstein GAN to address the small sample size problem in soft sensing. *Journal of Process Control*, 2022, **113**: 18–28
- 66 Pei Z Y, Jiang H K, Li X Q, Zhang J J, Liu S W. Data augmentation for rolling bearing fault diagnosis using an enhanced few-shot wasserstein auto-encoder with meta-learning. *Measurement Science and Technology*, 2021, **32**(8): Article No. 084007
- 67 Wang R G, Zhang S H, Chen Z Y, Li W H. Enhanced generative adversarial network for extremely imbalanced fault diagnosis of rotating machine. *Measurement*, 2021, **180**: Article No. 109467
- 68 Wang Y R, Sun G D, Jin Q. Imbalanced sample fault diagnosis of rotating machinery using conditional variational auto-encoder generative adversarial network. *Applied Soft Computing*, 2020, **92**: Article No. 106333
- 69 Liu S W, Jiang H K, Wu Z H, Li X Q. Data synthesis using deep feature enhanced generative adversarial networks for rolling bearing imbalanced fault diagnosis. *Mechanical Systems and Signal Processing*, 2022, **163**: Article No. 108139
- 70 Chawla N V, Bowyer K W, Hall L O, Kegelmeyer W P. SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 2002, **16**(1): 321–357
- 71 Mathew J, Pang C K, Luo M, Leong W H. Classification of imbalanced data by oversampling in kernel space of support vector machines. *IEEE Transactions on Neural Networks and Learning Systems*, 2018, **29**(9): 4065–4076
- 72 Maldonado S, López J, Vairetti C. An alternative SMOTE oversampling strategy for high-dimensional datasets. *Applied Soft Computing*, 2019, **76**: 380–389
- 73 Xie Hua, Chen Jun-Xing, Zhao Yu-Ming, Ding Qing, Zhang Pei. Knowledge acquisition method of power transformer condition assessment based on SMOTE and decision tree algorithm. *Electric Power Automation Equipment*, 2020, **40**(2): 137–142  
(谢桦, 陈俊星, 赵宇明, 丁庆, 张沛. 基于 SMOTE 和决策树算法的电力变压器状态评估知识获取方法. *电力自动化设备*, 2020, **40**(2): 137–142)
- 74 Liu Yun-Peng, He Jia-Hui, Xu Zi-Qiang, Wang Quan, Li Zhe, Gao Shu-Guo. Equalization method of power transformer fault sample based on SVM SMOTE. *High Voltage Engineering*, 2020, **46**(7): 2522–2529  
(刘云鹏, 和家慧, 许自强, 王权, 李哲, 高树国. 基于 SVM SMOTE 的电力变压器故障样本均衡化方法. *高电压技术*, 2020, **46**(7): 2522–2529)
- 75 Soltanzadeh P, Hashemzadeh M. RCSMOTE: Range-controlled synthetic minority over-sampling technique for handling the class imbalance problem. *Information Sciences*, 2021, **542**: 92–111
- 76 Li D C, Fang Y H. A non-linearly virtual sample generation technique using group discovery and parametric equations of hypersphere. *Expert Systems With Applications*, 2009, **36**(1): 844–851
- 77 Wang Shao-Jing, Ma Wen-Jia, Wang Feng-Hua, Cui Lü, Zhou Xing-Xing. Fault diagnosis of grounding grid based on virtual sample generation and probabilistic neural network. *High Voltage Apparatus*, 2020, **56**(6): 309–316  
(王劭菁, 马文嘉, 王丰华, 崔隼, 周行星. 基于虚拟样本生成技术与概率神经网络的接地网故障诊断. *高压电器*, 2020, **56**(6): 309–316)
- 78 Douzas G, Bacao F. Effective data generation for imbalanced learning using conditional generative adversarial networks. *Expert Systems With Applications*, 2018, **91**: 464–471
- 79 Huang Nan-Tian, Yang Xue-Hang, Cai Guo-Wei, Song Xing, Chen Qing-Zhu, Zhao Wen-Guang. A deep adversarial diagnosis method for wind turbine main bearing fault with imbalanced small sample scenarios. *Proceedings of the CSEE*, 2020, **40**(2): 563–573  
(黄南天, 杨学航, 蔡国伟, 宋星, 陈庆珠, 赵文广. 采用非平衡小样本数据的风机主轴轴承故障深度对抗诊断. *中国电机工程学报*, 2020, **40**(2): 563–573)
- 80 Li Z X, Zheng T S, Wang Y, Cao Z, Guo Z Q, Fu H Y. A novel method for imbalanced fault diagnosis of rotating machinery based on generative adversarial networks. *IEEE Transactions on Instrumentation and Measurement*, 2020, **70**: Article No. 3500417
- 81 Dixit S, Verma N K, Ghosh A K. Intelligent fault diagnosis of rotary machines: Conditional auxiliary classifier GAN coupled with meta learning using limited data. *IEEE Transactions on Instrumentation and Measurement*, 2021, **70**: Article No. 3517811
- 82 Yang G, Zhong Y, Yang L, Tao H, Li J Y, Du R X. Fault diagnosis of harmonic drive with imbalanced data using generative adversarial network. *IEEE Transactions on Instrumentation and Measurement*, 2021, **70**: Article No. 3519911
- 83 Zareapoor M, Shamsolmoali P, Yang J. Oversampling adversarial network for class-imbalanced fault diagnosis. *Mechanical Systems and Signal Processing*, 2021, **149**: Article No. 107175
- 84 Li Y B, Zou W T, Jiang L. Fault diagnosis of rotating machinery based on combination of wasserstein generative adversarial networks and long short term memory fully convolutional network. *Measurement*, 2022, **191**: Article No. 110826
- 85 Li M L, Zou D C, Luo S Y, Zhou Q, Cao L C, Liu H P. A new generative adversarial network based imbalanced fault diagnosis method. *Measurement*, 2022, **194**: Article No. 111045

- 86 Li Dong-Dong, Liu Yu-Hang, Zhao Yang, Zhao Yao. Fault diagnosis method of wind turbine planetary gearbox based on improved generative adversarial network. *Proceedings of the CSEE*, 2021, **41**(21): 7496–7506  
(李东东, 刘宇航, 赵阳, 赵耀. 基于改进生成对抗网络的风机行星齿轮箱故障诊断方法. *中国电机工程学报*, 2021, **41**(21): 7496–7506)
- 87 Hinton G E, Salakhutdinov R R. Reducing the dimensionality of data with neural networks. *Science*, 2006, **313**(5786): 504–507
- 88 Kingma D P, Welling M. Auto-encoding variational bayes. arXiv: 1312.6114, 2013.
- 89 Dai Jun, Wang Jun, Zhu Zhong-Kui, Shen Chang-Qing, Huang Wei-Guo. Anomaly detection of mechanical systems based on generative adversarial network and auto-encoder. *Chinese Journal of Scientific Instrument*, 2019, **40**(9): 16–26  
(戴俊, 王俊, 朱忠奎, 沈长青, 黄伟国. 基于生成对抗网络和自动编码器的机械系统异常检测. *仪器仪表学报*, 2019, **40**(9): 16–26)
- 90 Liu S W, Jiang H K, Wu Z H, Li X Q. Rolling bearing fault diagnosis using variational autoencoding generative adversarial networks with deep regret analysis. *Measurement*, 2021, **168**: Article No. 108371
- 91 Wang A Q, Qian Z, Pei Y, Jing B. A de-ambiguous condition monitoring scheme for wind turbines using least squares generative adversarial networks. *Renewable Energy*, 2022, **185**: 267–279
- 92 Liu Y P, Jiang H K, Wang Y F, Wu Z H, Liu S W. A conditional variational autoencoding generative adversarial networks with self-modulation for rolling bearing fault diagnosis. *Measurement*, 2022, **192**: Article No. 110888
- 93 Rathore M S, Harsha S P. Framework for imbalanced fault diagnosis of rolling bearing using autoencoding generative adversarial learning. *Journal of the Brazilian Society of Mechanical Sciences and Engineering*, 2023, **45**(1): Article No. 39
- 94 Huang C F. Principle of information diffusion. *Fuzzy Sets and Systems*, 1997, **91**(1): 69–90
- 95 Huang C F, Moraga C. A diffusion-neural-network for learning from small samples. *International Journal of Approximate Reasoning*, 2004, **35**(2): 137–161
- 96 Li D C, Wu C S, Tsai T I, Lina Y S. Using mega-trend-diffusion and artificial samples in small data set learning for early flexible manufacturing system scheduling knowledge. *Computers & Operations Research*, 2007, **34**(4): 966–982
- 97 Lin Y S, Li D C. The generalized-trend-diffusion modeling algorithm for small data sets in the early stages of manufacturing systems. *European Journal of Operational Research*, 2010, **207**(1): 121–130
- 98 Li D C, Chen C C, Chang C J, Lin W K. A tree-based-trend-diffusion prediction procedure for small sample sets in the early stages of manufacturing systems. *Expert Systems With Applications*, 2012, **39**(1): 1575–1581
- 99 Rahimi M R, Karimi H, Yousefi F. Prediction of carbon dioxide diffusivity in biodegradable polymers using diffusion neural network. *Heat and Mass Transfer*, 2012, **48**(8): 1357–1365
- 100 Zhu Bao, Chen Zhong-Sheng, Yu Le-An. A novel mega-trend-diffusion for small sample. *CIESC Journal*, 2016, **67**(3): 820–826  
(朱宝, 陈忠圣, 余乐安. 一种新颖的小样本整体趋势扩散技术. *化工学报*, 2016, **67**(3): 820–826)
- 101 Sivakumar J, Ramamurthy K, Radhakrishnan M, Won D. Synthetic sampling from small datasets: A modified mega-trend diffusion approach using k-nearest neighbors. *Knowledge-Based Systems*, 2022, **236**: Article No. 107687
- 102 Khamis N, Selamat H, Ismail F S. Improved optimization parameters prediction using the modified mega trend diffusion function for a small dataset problem. *Knowledge and Information Systems*, 2022, **64**(11): 3129–3149
- 103 Gao Ke-Xuan, Li Zhi-Gang, Xu Chang-Ming, Wang Qiao-Yun, Li Bo. Virtual sample establishment of Hybrid-MTD and its application in blood spectrum analysis. *Chinese Journal of Scientific Instrument*, 2019, **40**(8): 167–175  
(高克铤, 李志刚, 徐长明, 王巧云, 李博. 混合整体趋势扩散的虚拟样本构建及其血液光谱分析应用. *仪器仪表学报*, 2019, **40**(8): 167–175)
- 104 Chen Z S, Zhu B, He Y L, Yu L A. A PSO based virtual sample generation method for small sample sets: Applications to regression datasets. *Engineering Applications of Artificial Intelligence*, 2016, **59**: 236–243
- 105 Pawlak Z. Rough sets. *International Journal of Computer & Information Sciences*, 1982, **11**(5): 341–356
- 106 Hu Q W, Chakhar S, Siraj S, Labib A. Spare parts classification in industrial manufacturing using the dominance-based rough set approach. *European Journal of Operational Research*, 2017, **262**(3): 1136–1163
- 107 Chen W C, Chang N B, Chen J C. Rough set-based hybrid fuzzy-neural controller design for industrial wastewater treatment. *Water Research*, 2003, **37**(1): 95–107
- 108 Zhou P, Lu S W, Chai T Y. Data-driven soft-sensor modeling for product quality estimation using case-based reasoning and fuzzy-similarity rough sets. *IEEE Transactions on Automation Science and Engineering*, 2014, **11**(4): 992–1003
- 109 Li D C, Lin Y S. Using virtual sample generation to build up management knowledge in the early manufacturing stages. *European Journal of Operational Research*, 2006, **175**(1): 413–434
- 110 Li D C, Lin Y S. Learning management knowledge for manufacturing systems in the early stages using time series data. *European Journal of Operational Research*, 2008, **184**(1): 169–184
- 111 Li D C, Wen I H, Chen W C. A novel data transformation model for small data-set learning. *International Journal of Production Research*, 2016, **54**(24): 7453–7463
- 112 Li D C, Lin L S. A new approach to assess product lifetime performance for small data sets. *European Journal of Operational Research*, 2013, **230**(2): 290–298
- 113 Li D C, Lin L S, Chen C C, Yu W H. Using virtual samples to improve learning performance for small datasets with multimodal distributions. *Soft Computing*, 2019, **23**(22): 11883–11900
- 114 Liu S G, Zhang J, Xiang Y, Zhou W L. Fuzzy-based information decomposition for incomplete and imbalanced data learning. *IEEE Transactions on Fuzzy Systems*, 2017, **25**(6): 1476–1490
- 115 Ramentol E, Caballero Y, Bello R, Herrera F. SMOTE-RSB\*: A hybrid preprocessing approach based on oversampling and undersampling for high imbalanced data-sets using SMOTE and rough sets theory. *Knowledge and Information Systems*, 2012, **33**(2): 245–265
- 116 Hu Feng, Wang Lei, Zhou Yao. An oversampling method for imbalance data based on three-way decision model. *Acta Electronica Sinica*, 2018, **46**(1): 135–144  
(胡峰, 王蕾, 周耀. 基于三支决策的不平衡数据过采样方法. *电子学报*, 2018, **46**(1): 135–144)
- 117 Shen L J, Qian Q. A virtual sample generation algorithm supporting machine learning with a small-sample dataset: A case study for rubber materials. *Computational Materials Science*, 2022, **211**: Article No. 111475
- 118 Song H, Choi K K, Lee I, Zhao L, Lamb D. Adaptive virtual support vector machine for reliability analysis of high-dimensional problems. *Structural and Multidisciplinary Optimization*, 2013, **47**(4): 479–491
- 119 Li D C, Lin Y S. Generating information for small data sets with a multi-modal distribution. *Decision Support Systems*, 2014, **66**: 71–81
- 120 Nagarajan H P N, Mokhtarian H, Jafarian H, Dimassi S, Bakrani-Balani S, Hamed A, et al. Knowledge-based design of artificial neural network topology for additive manufacturing process modeling: A new approach and case study for fused de-

- position modeling. *Journal of Mechanical Design*, 2019, **141**(2): Article No. 021705
- 121 Link P, Poursanidis M, Schmid J, Zache R, von Kurnatowski M, Teicher U, et al. Capturing and incorporating expert knowledge into machine learning models for quality prediction in manufacturing. *Journal of Intelligent Manufacturing*, 2022, **33**(7): 2129–2142
- 122 Chen J Y, Geng Y X, Chen Z, Horrocks I, Pan J Z, Chen H J. Knowledge-aware zero-shot learning: Survey and perspective. In: Proceedings of the 30th International Joint Conference on Artificial Intelligence. Montreal, Canada: ijcai.org, 2021. 4366–4373
- 123 Zhuo Y, Ge Z Q. Auxiliary information-guided industrial data augmentation for any-shot fault learning and diagnosis. *IEEE Transactions on Industrial Informatics*, 2021, **17**(11): 7535–7545
- 124 Rethmeier N, Saxena V K, Augenstein I. TX-Ray: Quantifying and explaining model-knowledge transfer in (Un-)supervised NLP. In: Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence (UAI). AUAI Press, 2020. 440–449
- 125 Yao Z J, Zhao C H. FedTMI: Knowledge aided federated transfer learning for industrial missing data imputation. *Journal of Process Control*, 2022, **117**: 206–215
- 126 Feng L J, Zhao C H, Li X. Bias-eliminated semantic refinement for any-shot learning. *IEEE Transactions on Image Processing*, 2022, **31**: 2229–2244
- 127 Zhu Q X, Chen Z S, Zhang X H, Rajabifard A, Xu Y, Chen Y Q. Dealing with small sample size problems in process industry using virtual sample generation: A Kriging-based approach. *Soft Computing*, 2020, **24**(9): 6889–6902
- 128 Gong H F, Chen Z S, Zhu Q X, He Y L. A Monte Carlo and PSO based virtual sample generation method for enhancing the energy prediction and energy optimization on small data problem: An empirical study of petrochemical industries. *Applied Energy*, 2017, **197**: 405–415
- 129 Sang K H, Yin X Y, Zhang F C. Machine learning seismic reservoir prediction method based on virtual sample generation. *Petroleum Science*, 2021, **18**(6): 1662–1674
- 130 Li D C, Shi Q S, Li M D. Using an attribute conversion approach for sample generation to learn small data with highly uncertain features. *International Journal of Production Research*, 2018, **56**(14): 4954–4967
- 131 Lin Yue, Liu Ting-Zhang, Wang Zhe-He. Quantity optimization of virtual sample generation with two kinds of upper bound conditions. *Journal of Guangxi Normal University (Natural Science Edition)*, 2019, **37**(1): 142–148  
(林越, 刘廷章, 王哲河. 具有两类上限条件的虚拟样本生成数量优化. 广西师范大学学报(自然科学版), 2019, **37**(1): 142–148)
- 132 Yang Qing, Ye Yi-Xia, Wu Dong-Sheng, Liu Yi-Peng. Fault diagnosis for rolling bearings under variable working conditions based on ACGAN-DSAN. *Bearing*, 2023, (2): 97–104  
(杨青, 叶义霞, 吴东升, 刘伊鹏. 基于 ACGAN-DSAN 的变工况滚动轴承故障诊断. 轴承, 2023, (2): 97–104)
- 133 Li W X, Shang Z W, Gao M S, Qian S Q, Zhang B R, Zhang J. A novel deep autoencoder and hyperparametric adaptive learning for imbalance intelligent fault diagnosis of rotating machinery. *Engineering Applications of Artificial Intelligence*, 2021, **102**: Article No. 104279
- 134 Zhang X S, Zhuang Y, Wang W, Pedrycz W. Transfer boosting with synthetic instances for class imbalanced object recognition. *IEEE Transactions on Cybernetics*, 2018, **48**(1): 357–370
- 135 Liu J H, Ren Y F. A general transfer framework based on industrial process fault diagnosis under small samples. *IEEE Transactions on Industrial Informatics*, 2021, **17**(9): 6073–6083
- 136 Jia Xin, Gao Xin, Zhao Bing, Huang Zi-Jian, Ye Ping, Huang Xu. Intelligent electric meter fault classification based on sample migration and boundary enhancement in overlapping areas. *Power System Technology*, 2023, **47**(6): 2566–2582  
(贾欣, 高欣, 赵兵, 黄子健, 叶平, 黄旭. 基于样本迁移和交叠区边界增强的智能电表故障分类方法. 电网技术, 2023, **47**(6): 2566–2582)
- 137 Liao Yi-Fan, Wu Zhi-Gang. Critical sample generation method for static voltage stability based on transfer learning and Wasserstein generative adversarial network. *Power System Technology*, 2021, **45**(9): 3722–3728  
(廖一帆, 武志刚. 基于迁移学习与 Wasserstein 生成对抗网络的静态电压稳定临界样本生成方法. 电网技术, 2021, **45**(9): 3722–3728)
- 138 Lan Jian, Guo Qing-Lai, Zhou Yan-Zhen, Sun Hong-Bin. Generation of power system typical operation mode samples: A generation adversarial network and model-based transfer learning approach. *Proceedings of the CSEE*, 2022, **42**(8): 2889–2899  
(兰健, 郭庆来, 周艳真, 孙宏斌. 基于生成对抗网络和模型迁移的电力系统典型运行方式样本生成. 中国电机工程学报, 2022, **42**(8): 2889–2899)
- 139 Zhang L, Wei H, Lyu Z L, Wei H B, Li P J. A small-sample faulty line detection method based on generative adversarial networks. *Expert Systems With Applications*, 2020, **169**: Article No. 114378
- 140 Lee D, Kim K. An efficient method to determine sample size in oversampling based on classification complexity for imbalanced data. *Expert Systems With Applications*, 2021, **184**: Article No. 115442
- 141 Lin L S, Li D C, Chen H Y, Chiang Y C. An attribute extending method to improve learning performance for small datasets. *Neurocomputing*, 2018, **286**: 75–87
- 142 Lu Rong-Xiu, Lai Lu-Lu, Yang Hui, Zhu Jian-Yong. Prediction method of CePr/Nd component content based on hybrid virtual sample. *Control and Decision*, 2023, **38**(4): 1129–1136  
(陆荣秀, 赖路璐, 杨辉, 朱建勇. 基于混合虚拟样本生成的铈镧/钕组分含量预测. 控制与决策, 2023, **38**(4): 1129–1136)
- 143 Kang G Q, Wu L F, Guan Y, Peng Z. A virtual sample generation method based on differential evolution algorithm for overall trend of small sample data: Used for lithium-ion battery capacity degradation data. *IEEE Access*, 2019, **7**: 123255–123267
- 144 Napoli G, Xibilia M G. Soft sensor design for a topping process in the case of small datasets. *Computers & Chemical Engineering*, 2011, **35**(11): 2447–2456
- 145 Li Yan-Xia, Chai Yi, Hu You-Qiang, Yin Hong-Peng. Review of imbalanced data classification methods. *Control and Decision*, 2019, **34**(4): 673–688  
(李艳霞, 柴毅, 胡友强, 尹宏鹏. 不平衡数据分类方法综述. 控制与决策, 2019, **34**(4): 673–688)
- 146 Ma Bo, Cai Wei-Dong, Zhao Da-Li. Intelligent diagnosis method based on GAN sample generation technology. *Journal of Vibration and Shock*, 2020, **39**(18): 153–160  
(马波, 蔡伟东, 赵大力. 基于 GAN 样本生成技术的智能诊断方法. 振动与冲击, 2020, **39**(18): 153–160)
- 147 Li X Q, Jiang H K, Liu S W, Zhang J J, Xu J. A unified framework incorporating predictive generative denoising autoencoder and deep coral network for rolling bearing fault diagnosis with unbalanced data. *Measurement*, 2021, **178**: Article No. 109345
- 148 Pham M T, Kim J M, Kim C H. Rolling bearing fault diagnosis based on improved GAN and 2-D representation of acoustic emission signals. *IEEE Access*, 2022, **10**: 78056–78069
- 149 Yang Guang-You, Liu Lang, Xi Chen-Bo. Bearing fault diagnosis based on SA-ACGAN data generation model. *China Mechanical Engineering*, 2022, **33**(13): 1613–1621  
(杨光友, 刘浪, 习晨博. 自适应辅助分类器生成式对抗网络样本生成模型及轴承故障诊断. 中国机械工程, 2022, **33**(13): 1613–1621)
- 150 Pan T Y, Chen J L, Xie J S, Zhou Z T, He S L. Deep feature generating network: A new method for intelligent fault detection of mechanical systems under class imbalance. *IEEE Transactions on Industrial Informatics*, 2021, **17**(9): 6282–6293
- 151 Liu Yun-Peng, Xu Zi-Qiang, He Jia-Hui, Wang Quan, Gao Shu-

- Guo, Zhao Jun. Data augmentation method for power transformer fault diagnosis based on conditional Wasserstein generative adversarial network. *Power System Technology*, 2020, **44**(4): 1505–1513  
(刘云鹏, 许自强, 和家慧, 王权, 高树国, 赵军. 基于条件式 Wasserstein 生成对抗网络的电力变压器故障样本增强技术. *电网技术*, 2020, **44**(4): 1505–1513)
- 152 Liu J H, Qu F M, Hong X W, Zhang H G. A small-sample wind turbine fault detection method with synthetic fault data using generative adversarial nets. *IEEE Transactions on Industrial Informatics*, 2019, **15**(7): 3877–3888
- 153 Han T, Liu C, Yang W G, Jiang D X. A novel adversarial learning framework in deep convolutional neural network for intelligent diagnosis of mechanical faults. *Knowledge-Based Systems*, 2019, **165**: 474–487
- 154 Gao S Y, Wang X R, Miao X H, Su C W, Li Y B. ASM1D-GAN: An intelligent fault diagnosis method based on assembled 1D convolutional neural network and generative adversarial networks. *Journal of Signal Processing Systems*, 2019, **91**(10): 1237–1247
- 155 Olesen J F, Shaker H R. Predictive maintenance within combined heat and power plants based on a novel virtual sample generation method. *Energy Conversion and Management*, 2021, **227**: Article No. 113621
- 156 Wu Xiao-Dong, Xiong Wei-Li. Fault detection method and its application using GAN with an encoded input. *CAAI Transactions on Intelligent Systems*, 2022, **17**(3): 496–505  
(吴晓东, 熊伟丽. 采用编码输入的生成对抗网络故障检测方法及应用. *智能系统学报*, 2022, **17**(3): 496–505)
- 157 Xu Y, Zhao Y, Ke W, He Y L, Zhu Q X, Zhang Y, et al. A multi-fault diagnosis method based on improved SMOTE for class-imbalanced data. *The Canadian Journal of Chemical Engineering*, 2023, **101**(4): 1986–2001
- 158 Karras T, Aittala M, Hellsten J, Laine S, Lehtinen J, Aila T. Training generative adversarial networks with limited data. In: *Proceedings of the 34th International Conference on Neural Information Processing Systems*. Vancouver, Canada: Curran Associates Inc., 2020. 12104–12114
- 159 Zhuang Fu-Zhen, Luo Ping, He Qing, Shi Zhong-Zhi. Survey on transfer learning research. *Journal of Software*, 2015, **26**(1): 26–39  
(庄福振, 罗平, 何清, 史忠植. 迁移学习研究进展. *软件学报*, 2015, **26**(1): 26–39)
- 160 Pang Jing-Yue, Zhao Guang-Quan. Digital twin-driven multi-algorithms adaptive selection for fault detection of space power system. *Journal of Electronic Measurement and Instrumentation*, 2022, **36**(6): 91–99  
(庞景月, 赵光权. 数字孪生驱动多算法自适应选择的空间电源系统故障检测. *电子测量与仪器学报*, 2022, **36**(6): 91–99)
- 161 Zhang Xu-Hui, Ju Jia-Shan, Yang Wen-Juan, Lv Xin-Yuan. Predictive maintenance system for complex mining equipment based on digital twin. *Chinese Journal of Engineering Design*, 2022, **29**(5): 643–650  
(张旭辉, 鞠佳杉, 杨文娟, 吕欣媛. 基于数字孪生的复杂矿用设备预测性维护系统. *工程设计学报*, 2022, **29**(5): 643–650)
- 162 Sun Zi-Jian, Tang Jian, Qiao Jun-Fei. Semi-supervised concept drift detection method by combining sample output space and feature space with its application. *Acta Automatica Sinica*, 2022, **48**(5): 1259–1272  
(孙子健, 汤健, 乔俊飞. 联合样本输出与特征空间的半监督概念漂移检测法及其应用. *自动化学报*, 2022, **48**(5): 1259–1272)
- 163 Xu Wen, Tang Jian, Xia Heng, Qiao Jun-Fei. Soft sensor of dioxin emission concentration based on Bagging semi-supervised deep forest regression. *Chinese Journal of Scientific Instrument*, 2022, **43**(6): 251–259  
(徐雯, 汤健, 夏恒, 乔俊飞. 基于 Bagging 半监督深度森林回归的二噁英排放浓度软测量. *仪器仪表学报*, 2022, **43**(6): 251–259)

- 164 Qiao Jun-Fei, Sun Zi-Jian, Tang Jian. Overview of concept drift detection for industrial process soft sensor modeling. *Control Theory & Applications*, 2021, **38**(8): 1159–1174  
(乔俊飞, 孙子健, 汤健. 面向工业过程软测量建模的概念漂移检测综述. *控制理论与应用*, 2021, **38**(8): 1159–1174)



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

E-mail: freeflytang@bjut.edu.cn

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



崔璨麟 北京工业大学信息学部硕士研究生. 主要研究方向为城市固废焚烧过程风险预警, 虚拟样本生成.

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

(CUI Can-Lin Master student at the Faculty of Information Technology, Beijing University of Technology. His research interest covers risk warning of municipal solid waste incineration process and virtual sample generation.)



夏恒 北京工业大学信息学部博士研究生. 主要研究方向为树结构深/宽度学习结构设计与优化, 城市固废焚烧过程二噁英排放预测.

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

(XIA Heng Ph.D. candidate at the Faculty of Information Technology, Beijing University of Technology. His research interest covers structure design and optimization of tree-structured deep/broad learning and dioxin emission prediction of the municipal solid waste incineration process.)



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

E-mail: junfeiq@bjut.edu.cn

(QIAO Jun-Fei Professor at the Faculty of Information Technology, Beijing University of Technology.

His research interest covers intelligent control of wastewater treatment process, and structure design and optimization of neural networks.)