DOI:10.19651/j.cnki.emt.2416647

融入噪声的监督增强网络用于小样本数据增强*

郭小萍 赵霄丰 李 元
(沈阳化工大学信息工程学院 沈阳 110142)

摘 要: 在复杂的工业过程中,由于关键变量难以测量,过程数据具有不平衡和不完整的特点,导致软测量性能下降。为了解决这一问题,提出一种融入噪声的监督增强自编码器虚拟样本生成方法。首先,为了加强输入与输出的映射关系并保证特征信息的完整性,该方法在自编码器的编码部分添加增强层,解码部分引入标签信息进行有监督约束训练。为了增加虚拟样本的多样性,在监督增强自编码器隐藏层提取的特征中加入高斯噪声。将生成的虚拟样本与原始小样本相结合,增强软测量模型的性能。与传统的虚拟样本生成方法不同,所提的 NISEAE-VSG 模型可以同时生成输入输出虚拟样本。为了验证所提方法的有效性,使用火力发电和聚乙烯过程的数据集进行仿真验证。仿真结果表明,所提方法生成的虚拟样本优于其他虚拟样本生成方法,可有效提高软测量建模精度。

关键词: 软测量;虚拟样本;小样本;数据增强;监督增强自编码器
中图分类号: TN081 **文献标识码:** A **国家标准学科分类代码:** 510.80

Supervised enhancement network incorporating noise for small sample data augmentation

Guo Xiaoping Zhao Xiaofeng Li Yuan
(College of Information Engineering, Shenyang University of Chemical Technology, Shenyang 110142, China)

Abstract: In the complex industrial processes, the process data have the characteristics of imbalance and are incomplete due to the difficult-to-measure key variables, leading to the performance degradation of soft sensors. In order to deal with this problem, a novel noise injection supervised enhanced autoencoder virtual sample generation method is proposed. Firstly, in order to strengthen the mapping relationship between input and output and ensure the integrity of feature information, this method adds an enhancement layer to the encoding part of the autoencoder, and introduces label information for supervised constraint training in the decoding part. In order to increase the diversity of virtual samples, Gaussian noise is added to the features extracted from the hidden layer of the supervised enhanced autoencoder. Combine the generated virtual samples with the original small samples to enhance the performance of the soft sensing model. Unlike traditional virtual sample generation methods, the proposed NISEAE-VSG model can simultaneously generate useful input-output virtual samples. To verify the effectiveness of the proposed method, simulations were conducted using datasets of thermal power generation and polyethylene processes. The simulation results show that the proposed method generates virtual samples that are superior to other virtual sample generation methods and can effectively improve the accuracy of soft sensing modeling.

Keywords: soft sensor;virtual samples;small samples;data augmentation;supervised enhanced autoencoder

0 引 言

随着工业生产规模的不断扩大,影响生产过程的因素越来越多,如何获得充分而有价值的过程数据仍然是建立准确的软测量模型的关键[1]。影响这一关键点的因素有两个:首先,由于数据信息的不确定性以及噪声和缺失值的存在,从大数据中提取有用信息变得非常困难[2];其次,收集一些难以测量的关键变量往往需要较高的测量成本[3]。如果直接使用这些信息不足、分布未知、表示不佳的样本来构建软测量,将会大大降低模型的精度。一般来说,上述问题可以定义为小样本问题[4]。小样本问题的出现对建立一个准确的模型是一个很大的挑战。

收稿日期:2024-08-10
* 基金项目:国家自然科学基金(62273242)、辽宁省教育厅科学研究一般项目(LJ2020021)资助

为了处理小样本问题，已经提出了许多先进技术。这些技术大致分为 3 类：第一类是基于灰色的方法，二是基于降维技术的特征提取方法，最后，也是最流行的一种方法，基于虚拟样本生成(virtual sample generation，VSG)的方法。基于 VSG 的方法可以根据原始小样本生成虚拟样本，各种实验表明，加入新生成的虚拟样本可以提高模型的精度[5]。

近年来，基于 VSG 的方法受到了相关研究者的广泛关注。基于信息扩散的 VSG 方法可以利用信息扩散原理，利用模糊理论推导出原始小样本的扩散函数，生成一定范围内的样本[6]。典型的基于信息扩散的 VSG 包括大趋势扩散(mega-trend-diffusion，MTD)[7]和基于树的趋势扩散(tree-based trend diffusion，TTD)[8]方法。然而，基于信息扩散的 VSG 的不便之处在于扩散函数和参数往往难以确定。此外，使用基于信息扩散的 VSG 生成的虚拟样本缺乏输入和输出变量之间的相关性。为了解决这些问题，He 等[9]利用极限学习机(extreme learning machine，ELM)的隐藏表示生成虚拟样本，克服了参数不确定的问题。为了减少虚拟样本的引入导致误差积累的效应，王丹丹等[10]提出了一种基于多目标粒子群(particle swarm optimization，PSO)混合优化的虚拟样本生成(hybrid optimization with Multi-objective PSO，MOPSO)方法。该方法可以保证虚拟样本分布与小样本分布相似，获得大量高质量的虚拟样本并确定最佳虚拟样本数量。此外，生成式对抗网络(generative adversarial networks，GAN)等特征提取的 VSG 在生成高维数据和增加样本多样性方面得到了广泛的应用[11]。然而，上述的 VSG 方法通常是先产生虚拟样本的输入，然后再获得虚拟样本的输出。忽略虚拟样本输入和输出之间的相关性。因此，基于这些虚拟样本的软测量模型的性能仍然有限。根据自动编码(autoencoder，AE)理论，AE 是产生新的输入和输出的有效的非线性工具。Xie 等[12]提出了一种监督变分自编码器(supervised variational AE，SVAE)来处理缺失数据，以提高软测量模型的精度。但是，如果分别对输入和输出进行 AE 重构，则会破坏重构输入和输出之间的映射关系，对建模精度造成不良影响。

为进一步解决小样本问题，建立性能优越的软测量模型，本文提出了一种融入噪声的监督增强自编码器虚拟样本生成方法(noise injection supervised enhanced autoencoder virtual sample generation method，NISEAE-VSG)。与传统 AE 模型相比，监督增强 AE 模型考虑了原始输入和输出之间的映射关系。通过 SEAE，可以同时生成虚拟输入和输出样本，并保持输入和输出之间的原始映射关系。为了增加虚拟样本的多样性，将高斯噪声注入 SEAE 隐藏层提取的特征中。所提出方法可以生成可行且有用的虚拟样本，以提高软测量模型的性能。为了验证所提出的 NISEAE-VSG 的优越有效性和可靠可行性，以火力发电和聚乙烯生产过程为例，构建了基于 NISEAE-VSG 的 ELM 软测量模型。

1　基本理论

VSG 方法生成的虚拟样本与原始小样本结合，增强了模型的性能。图 1 给出了小样本与虚拟样本的关系，在生成虚拟样本的过程中，会生成许多好的虚拟样本(橙色菱形)，也会产生一些不好的虚拟样本(黑点)。好的虚拟样本要保存，不好的虚拟样本要丢弃。

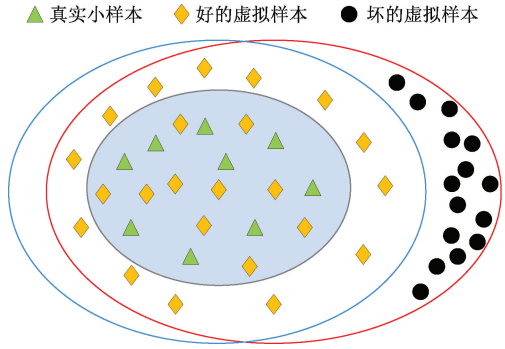

图 1　小样本和虚拟样本说明

Fig. 1　Illustration of small samples and virtual samples

1.1　堆叠自编码器

堆叠自编码器(stacked autoencoder，SAE)是一种典型的深度学习模型，由多个 AE 逐层叠加而成，SAE 的结构如图 2 所示。AE 由编码和解码部分组成。在编码部分，输入数据通过非线性映射传输到隐藏层，编码器的输出可以看作是输入因子的提取特征，如式(1)所示。解码器的目的是根据编码器的表示重构输入，使解码器的输出近似等于输入，如式(2)所示。

$$h = f(Wx + b) \tag{1}$$

$$\tilde{x} = g(W'x + b') \tag{2}$$

其中，输入 $x = [x_1, x_2, \cdots, x_{dx}]$；重构输出 $\tilde{x} = [\tilde{x}_1, \tilde{x}_2, \cdots, \tilde{x}_{dx}]$；$h$ 是提取的特征；$\{W, b\}$ 和 $\{W', b'\}$ 分别表示编码与解码部分的权重和偏差矩阵。自编码器的训练过程可表示为：

$$J(\theta) = \frac{1}{2N}\sum_{i=1}^{N} \| \tilde{x}_i - x_i \|^2 \tag{3}$$

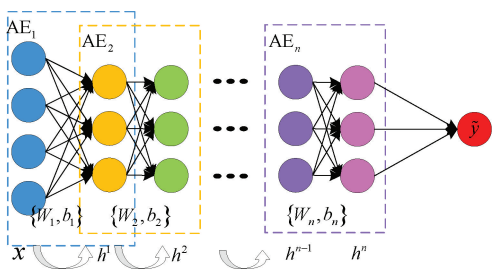

图 2　SAE 结构

Fig. 2　Structure of SAE

堆叠多个 AE 构成 SAE，提取数据的深层特征。在微

调过程中通过最小化质量预测误差式(4)优化网络参数。

$$J(\theta)=\frac{1}{2N}\sum_{i=1}^{N}\|\tilde{y}_i-y_i\|^2 \tag{4}$$

1.2 极限学习机

ELM 结构如图 3 所示,是一种三层前馈拓扑结构,具有收敛速度快,精度高的特点[13]。在 ELM 中,不需要提前调整参数。输入层和隐藏层之间的连接权 W_{mL} 是随机分配的。输出权值 β_{Lo} 可以通过最小二乘解来计算。假设小样本为 $S=\{(x_i,y_i),i=1,2,\cdots,M\}$,其中,输入向量 $x_i=[x_{i1},x_{i2},\cdots,x_{im}]^{\mathrm{T}}\in\mathbb{R}^m$,且 $y_i=[y_{i1},y_{i2},\cdots,y_{im}]^{\mathrm{T}}\in\mathbb{R}^o$ 表示输出向量。ELM 的隐藏节点是数是 L,ELM 的输出可以计算为:

$$y_i=\sum_{j}^{L}\beta_j f(w_j\cdot x_i+b_j),i=1,2,\cdots,M \tag{5}$$

其中,输入层与隐藏层之间的连接权值为 $W_j=[w_{j1},w_{j2},\cdots,w_{jm}]^{\mathrm{T}}$,$\beta_j=[\beta_{j1},\beta_{j2},\cdots,\beta_{jm}]^{\mathrm{T}}$ 表示输出权值,b_j 表示隐层偏置向量,$f(\cdot)$ 表示非线性激活函数。

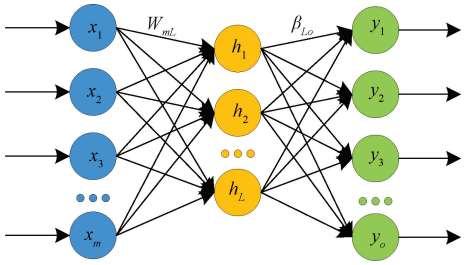

图 3　ELM 算法结构

Fig. 3　Basic ELM algorithm

利用最小二乘法,可以得到输出权值 β 为:

$$\beta=\boldsymbol{F}^{\dagger}\cdot y \tag{6}$$

其中,$\boldsymbol{F}^{\dagger}$ 表示 $\boldsymbol{F}$ 的穆尔-彭罗斯广义逆矩阵。

2 NISEAE-VSG 模型软测量方法

2.1 NISEAE-VSG 模型的构建

本文所提的 NISEAE-VSG 如图 4 所示,其中 e 是前一层 SEAE 提取的隐藏层表示 h,作为当前 SEAE 的增强层。SEAE 重构原始输入的同时保留特征提取过程中输入与输出之间的映射关系,增强层的引入保留了提取特征信息的完整性,减少随着网络层数的增加信息损失的逐层累积。与原始 AE 相比,SEAE 在输出层中引入目标输出变量,减少了不相关的隐藏信息。因此,SEAE 可以增强输入数据和输出数据之间的内部相关性。此外,为了提高生成虚拟样本信息的多样性,将高斯噪声 G 随机注入 SEAE 隐藏层的节点中,构成 NISEAE,将 NISEAE 的解码重构部分通过权值矩阵计算生成的虚拟样本 $\tilde{x}v$ 和 $\tilde{y}v$,填补小样本的数据缺口。

将归一化的数据集 $[\boldsymbol{X},\boldsymbol{Y}]$ 作为 SEAE 的输入和输出。在编码器部分,输入数据 $\boldsymbol{X}$ 通过式(7)转换为 h 表示的隐藏表示。在解码器部分,SEAE 的期望输出由原始输

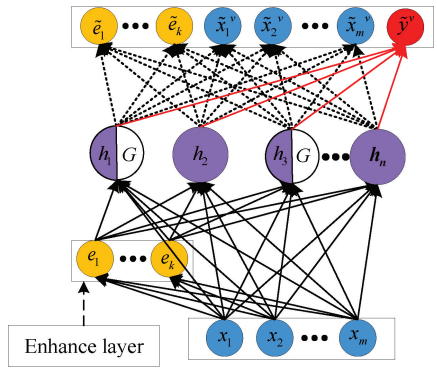

图 4　NISEAE-VSG 方法结构

Fig. 4　Structure of NISEAE-VSG model

入变量 $\boldsymbol{X}$ 和目标输出变量 $\boldsymbol{Y}$ 组成,SEAE 的实际输出如下:

$$h=f(Wx+b) \tag{7}$$

$$[\tilde{x},\tilde{y}]=f(\tilde{W}h+\tilde{b}) \tag{8}$$

最小化 SEAE 的损失函数进行训练。SEAE 的损失函数定义为:

$$J(\theta)=\frac{1}{2N}\sum_{i=1}^{N}\|\tilde{x}_i-x_i\|^2+\frac{1}{2N}\sum_{j=2}^{k}\sum_{i}^{N}\|h_i^j-h_i^{(j-1)\cdot k}\|^2+\frac{1}{2N}\sum_{i=1}^{N}\|\tilde{y}_i-y_i\| \tag{9}$$

从式(9)中可以看出,重构误差与增强层被引入到 SEAE 的损失函数中,说明 SEAE 的学习是有监督的,SEAE 可以学习输入和输出之间的映射关系,并确保提取特征信息的完整性。因此,SEAE 可以直接重建原始输入和输出。为了生成更丰富、更多样化的虚拟样本,在 SEAE 隐藏层中随机注入高斯噪声 G。假设输入层与隐藏层之间的连接权为 $\boldsymbol{W}_j=[w_{j1},w_{j2},\cdots,w_{jn}]^{\mathrm{T}}$;隐层偏置向量 $\boldsymbol{b}=[b_1,b_2,\cdots,b_n]$,$G$ 定义为 $G=\mathrm{wgn}(N,1,0)$,是一个均值为 0,方差为 1,信号强度为 0 dBW 的高斯噪声。首先,随机选择列 $[c_1,c_2,\cdots,c_s]$,$c_s\leqslant n$,注入 G;然后,注入 G 后的式(7)改写成:

$$h'=[\boldsymbol{f}_1,\cdots,(f_{c_1}+G^{c_1}),\cdots,\boldsymbol{f}_j,\cdots,(f_{c_q}+G^{c_q}),\cdots,\boldsymbol{f}_n] \tag{10}$$

其中,$\boldsymbol{f}_j$ 是矩阵的缩写:

$$\boldsymbol{f}_j=\begin{bmatrix}f(w_j\cdot x_1+b_j)\\ \cdots\\ f(w_j\cdot x_M+b_j)\end{bmatrix},j=1,2,\cdots,n \tag{11}$$

$\boldsymbol{G}^{c_q}$ 是高斯噪声矩阵的缩写:

$$\boldsymbol{G}^{c_q}=\begin{bmatrix}\gamma\cdot G_1\\ \vdots\\ \gamma\cdot G_M\end{bmatrix}^{c_q}_{M\times 1},q=1,2,\cdots,s \tag{12}$$

若 $j=c_q$,则注入高斯噪声 G 的新 c_q 列可表示为 $f_{c_q}+G^{c_q}$。γ 是避免样本被高斯噪声 G 完全覆盖的比例因子。因此,设 G 的值为 $\gamma\cdot G\ll f(w\cdot x+\boldsymbol{b})$。

由于隐藏层和输出层之间的连接权重 $\tilde{W}$ 和偏置向量 $\tilde{b}$ 已经进行了预训练，因此 NISEAE 模型的原始输入和输出生成的虚拟样本可以表示为：

$$[\tilde{x}^v, \tilde{y}v] = f(\tilde{W}h' + \tilde{b}) \tag{13}$$

2.2　NISEAE-VSG 模型软测量方法

本文提出的 NISEAE-VSG 模型可以通过隐藏层特征信息生成与输入和输出相关联的虚拟样本。为了提高模型的性能，将虚拟样本添加到训练集中。本文采用 ELM 建立软测量模型。本文分为两部分来说明虚拟样本对软测量模型精度的影响：一是利用原始小样本构建 ELM 软测量模型；另一种是在小样本中添加虚拟样本。下面详细介绍了虚拟样本的生成过程以及对软测量模型进行 NISEAE-VSG 有效性验证的过程。

1）将小样本数据的输入 $\boldsymbol{X}$ 和输出 $\boldsymbol{Y}$ 归一化到[0,1]的范围内。将数据集 $[\boldsymbol{X},\boldsymbol{Y}]$ 划分为训练数据集 $[\boldsymbol{X}_{train},\boldsymbol{Y}_{train}]$ 和测试数据集 $[\boldsymbol{X}_{test},\boldsymbol{Y}_{test}]$。设置隐藏节点数为 n，确定比例因子 γ 值。

2）通过 BP 算法使用 $[\boldsymbol{X}_{train},\boldsymbol{Y}_{train}]$ 训练 NISEAE；计算 $\{W,b,\tilde{W},\tilde{b}\}$，得到 $h = \{h_1,h_2,\cdots,h_n\}$。

3）生成高斯噪声 G，随机选取一部分隐藏节点；将 G 注入选定的隐藏节点，带有 G 的隐藏特征表示计算为 h'。

4）使用预训练好的参数集 $\{\tilde{W},\tilde{b}\}$ 重新计算 NISEAE 的输出；通过式(13)可以获得虚拟样本 $[\boldsymbol{X}^v_{train},\boldsymbol{Y}^v_{train}]$。

5）保存隐藏表示 h 和参数集 $\{W,b,\tilde{W},\tilde{b}\}$，重复步骤 3)和 4)，生成更多的虚拟样本。

6）使用 $[\boldsymbol{X}_{train},\boldsymbol{Y}_{train}]$ 和 $[\boldsymbol{X}^v_{train},\boldsymbol{Y}^v_{train}]$ 构建 ELM 软测量模型。采用 $[\boldsymbol{X}_{test},\boldsymbol{Y}_{test}]$ 分别对模型在有无虚拟样本情况下的性能进行验证。

7）通过计算平均相对误差(average relative error，ARE)和确定系数(coefficient of determination，R^2)来评价 NISEAE-VSG 的可行性和有效性。用误差改善率(error improving rate，EIR_ARE)来说明软测量精度的提高。ARE，R^2 和 EIR_ARE 分别在式(14)～(16)中计算。

$$ARE = \frac{1}{N}\sum_{i}^{N}\left|\frac{\tilde{\boldsymbol{Y}}_{i,test} - \boldsymbol{Y}_{i,test}}{\boldsymbol{Y}_{i,test}}\right| \tag{14}$$

$$R^2 = \frac{\sum_{i=1}^{N}(\tilde{\boldsymbol{Y}}_{i,test} - \boldsymbol{Y}_{i,test})^2}{\sum_{i=1}^{N}(\boldsymbol{Y}_{i,test} - \bar{\boldsymbol{Y}}_{i,test})^2} \tag{15}$$

$$EIR\_ARE = \frac{ARE_{before} - ARE_{after}}{ARE_{before}} \times 100\% \tag{16}$$

8）为了确定新获得的虚拟数据的质量，分别计算样本的平均值 M 和方差 S^2：

$$M = \frac{\sum_{i=1}^{m} a_i}{m} \tag{17}$$

$$S^2 = \frac{\sum_{i=1}^{m}(a_i - M)^2}{m} \tag{18}$$

其中，m 为样本总数，a_i 为第 i 个样本的变量。

3　案例验证

使用工业聚乙烯[14]和火力发电[15]过程的两个数据集验证所提 NISEAE-VSG 模型的有效性。并与大趋势扩散(MTD)，基于树的趋势扩散(TTD)，变分自编码器(VAE)，生成对抗网络(GAN)和融入噪声的目标相关自编码器(noise injection target-relevant autoencoder，NITAE)[16]对比验证。

3.1　工业聚乙烯案例

聚乙烯的工业生产过程中有 3 个主要反应器。反应器的关键变量熔体指数(melt index，MI)。本文选择第一级反应器的 12 个辅助变量作为模型的输入数据，MI 含量作为质量变量。从文献[14]中选择 250 个样本建模，其中前 200 个样本用于模型训练，其余 50 个用于模型测试。

在训练过程中，学习率 0.03，迭代次数 500 次。比例因子 0.01。为了说明新生成的虚拟样本的质量，表 1 给出了原始样本和虚拟样本的均值和方差。从表 1 可以明显看出，虚拟变量的均值和方差与原始变量的均值和方差结果非常相差很小，说明生成的虚拟样本是合理的。

表 1　乙烯数据小样本质量分析

Table 1　Small sample quality analysis of ethylene dataset

变量	M		S^2	
	O	V	O	V
x_1	0.267 3	0.262 0	0.042 8	0.044 1
x_2	0.341 5	0.335 3	0.069 8	0.067 8
x_3	0.289 4	0.296 8	0.049 8	0.044 9
x_4	0.601 5	0.604 1	0.020 1	0.022 3
x_5	0.369 0	0.370 2	0.054 4	0.056 9
x_6	0.496 6	0.498 3	0.032 3	0.034 6
x_7	0.410 3	0.416 3	0.034 7	0.034 4
x_8	0.327 3	0.327 9	0.012 6	0.013 0
x_9	0.426 4	0.433 5	0.037 6	0.037 5
x_{10}	0.308 0	0.316 4	0.041 1	0.044 2
x_{11}	0.487 0	0.481 3	0.031 6	0.033 0
x_{12}	0.362 2	0.365 7	0.123 0	0.121 6
y	0.331 4	0.319 3	0.090 9	0.077 9

然后，构建基于 ELM 的软测量来评估所提出的 NISEAE-VSG 的有效性。为了探索虚拟样本对 ELM 性能的影响，在训练集中分别添加不同大小的虚拟样本。只改变训练集的大小，而测试集保持不变，以验证模型的性能。添加到训练集中的虚拟样本的数量分别为 20、60、100、120 和 200。聚乙烯软测量模型性能分析如表 2 所示。由表 2 可知，未添加虚拟样本的 ELM 软测量 ARE 值大于添加虚

拟样本的 ELM 软测量 ARE 值。显然,通过添加不同大小的虚拟样本,软测量模型的 ARE 值都有不同程度的提高,改进后的 EIR_ARE 最大可达到 41.35%,说明虚拟样本是有效的,对精度的提高有积极的影响。

表 2 NISEAE-VSG 模型性能说明

Table 2 Performance description of NISEAE-VSG model

样本数	ARE	EIR_ARE
200	0.533 9	0
200+20	0.453 7	0.150 2
200+60	0.405 5	0.240 4
200+100	0.313 2	0.413 5
200+120	0.328 2	0.385 2
200+200	0.375 1	0.297 4

为了进一步证明所提出模型的优越性,与 MTD、TTD、VAEVSG、NITAEVSG、GANVSG 等 5 种 VSG 方法进行了对比实验,其中 VSG 表示虚拟样本生成。表 3 给出了所提方法与其他 5 种 VSG 方法的 ARE 和 R^2 指标。从表 3 可以看出,NISEAE-VSG 的建模精度最高,最小 $ARE=0.313\ 2$,最大 $R^2=8\ 901$。MTD 和 TTD 方法可以通过扩散函数生成新的虚拟样本,作为原始样本的信息补充。然而,扩散函数和参数难以确定,导致虚拟样本质量较差。因此,MTD 和 TTD 方法结果最差。同样,VAEVSG,GANVSG 和 NITAEVSG 方法生成的输入虽然可以遵循原始样本的条件分布,但生成的虚拟输入与输出之间的映射关系不足,导致建模性能不如所提出的 NISEAE-VSG。总的来说,所提方法生成的虚拟样本对于提高软测量建模精度是有效的。

表 3 加入不同的数量的虚拟样本比较不同方法的 ARE 和 R^2

Table 3 Comparing ARE and R^2 of different methods by adding different numbers of virtual samples

方法	虚拟样本数量									
	20		60		100		120		200	
	ARE	R^2	ARE	R^2	ARE	R^2	ARE	R^2	ARE	R^2
MTD	0.690 9	0.596 6	0.638 3	0.652 8	0.682 9	0.586 3	0.698 0	0.542 3	0.723 1	0.523 7
TTD	0.681 1	0.575 0	0.663 1	0.592 7	0.618 4	0.672 1	0.652 7	0.607 3	0.714 7	0.510 4
VAEVSG	0.538 4	0.699 7	0.493 6	0.748 1	0.449 2	0.776 4	0.430 2	0.793 6	0.607 4	0.630 3
GANVSG	0.565 5	0.673 2	0.526 7	0.709 4	0.498 7	0.739 4	0.569 8	0.680 6	0.590 4	0.658 6
NITAEVSG	0.525 4	0.706 0	0.473 9	0.740 2	0.439 4	0.796 8	0.399 6	0.820 1	0.440 9	0.776 8
NISEAEVSG	0.453 7	0.730 1	0.405 5	0.793 7	0.313 2	0.890 1	0.328 2	0.868 1	0.375 1	0.803 1

此外,当 6 种 VSG 方法的软测量建模达到最高精度时,添加和未添加虚拟样本的测试输出如图 5 所示。由图 5 可知,没有虚拟样本的情况下,模型性能并不理想。在加入虚拟样本后,可以提高模型的精度。与其他 5 种 VSG 方法相比,NISEAE-VSG 的测试输出(蓝色线)更接近预期输出(红色线)。6 种方法的预测误差箱形图如图 6 所示,与其他模型相比,所提方法可以实现更窄的误差区

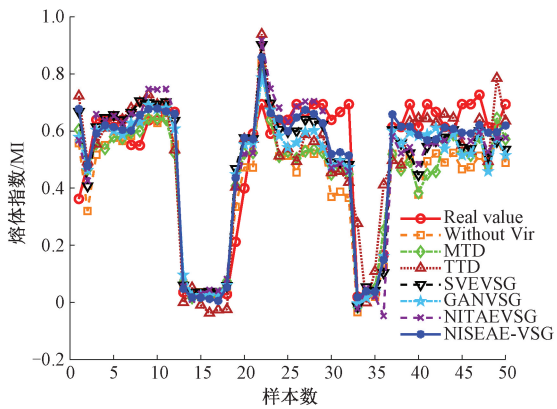

图 5 乙烯添加虚拟样品前后的测试输出

Fig. 5 Test output before and after adding virtual samples

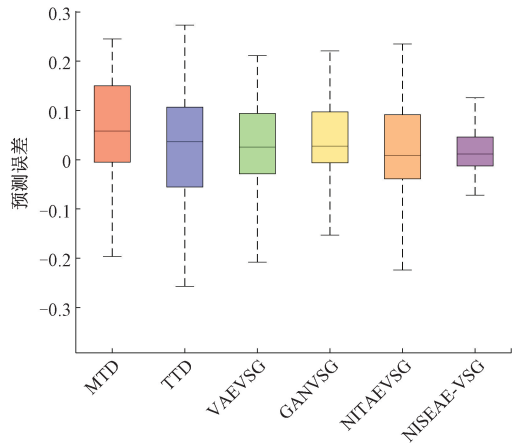

图 6 乙烯测试输出误差的箱形图

Fig. 6 Box plot of output error in ethylene testing

间,中值线(黑线)更接近于 0。另外,6 种 VSG 方法测试输出的散点图如图 7 所示。与其他模型的数据点相比,NISEAE-VSG 的数据点(蓝色)更接近于预期的输出值(黑线)。验证,所提 NISEAE-VSG 模型的有效性和优越性。

3.2 火力发电案例

火力发电是一种常用的工业发电方式。其中,蒸汽流

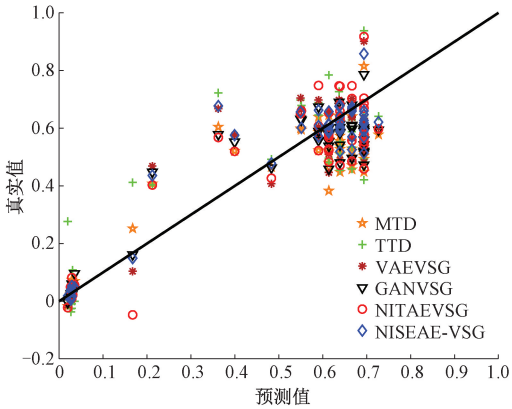

图 7　乙烯 6 种 VSG 法测试输出的散点图

Fig. 7　Scatter plots of 6 VSG test outputs for ethylene

量(steam flow，SF)是影响发电效率的关键指标。影响 SF 的因素共有 38 个，包括锅炉的氧气含量、可燃物的进料速率、取水量等。质量变量是 SF。共收集了 500 个样本，前 400 个样本用作训练集，其余样本用作测试集。

模型学习率 0.01，迭代次数 500。比例因子 0.01。表 4 给出了原始样本和虚拟样本的均值和方差。可以看出虚拟变量的方差与原始变量的方差非常接近，说明生成的虚拟样本是合理可行的。为了进一步验证 NISEAE-VSG 对软测量模型性能提升的影响，训练集中添加不同数量的虚拟样本，分别为 40、80、160、240 和 320。表 5 给出了基于 ELM 的 NISEAE-VSG 的评价指标 ARE 和 EIR_ARE 的值。由表 5 可知，EIR_ARE 的最大值为 63.36%，通过增加虚拟样本可以提高软测量模型的性能。

为了揭示 NISEAE-VSG 在处理小样本问题方面的优势，选择 MTD、TTD、VAEVSG、GANVSG、NITAEVSG 这 5 种 VSG 方法进行对比研究。表 6 列出了 6 种 VSG 方法的 ARE 和 R^2 指标。从表 6 可以看出，与其他 5 种 VSG 方法相比，所提方法的 ARE 值最小，为 0.125 5，R^2 值最大，为 0.941 7。受扩散函数和参数选择困难的影响，MTD 和 TTD 方法的软测量建模效果最差。VAEVSG 和 GANVSG 是先产生虚拟样本的输入，然后再获得虚拟样本的输出，忽略虚拟样本输入和输出之间的相关性。尽管，NITAEVSG 能同时生成输入与输出的虚拟样本，但忽视了生成的虚拟样本信息的完整性，建模性能不如所提出的 NISEAE-VSG。

表 4　火力发电数据小样本质量分析

Table 4　Quality analysis of small sample of thermal power plant data

变量	M		S^2	
	O	V	O	V
x_1	0.708 3	0.698 1	0.011 5	0.013 3
x_2	0.743 7	0.755 3	0.011 1	0.010 6
x_3	0.641 6	0.640 1	0.010 6	0.013 7
x_4	0.639 0	0.631 9	0.016 1	0.018 1
x_5	0.643 7	0.653 8	0.005 9	0.006 2
⋮	⋮	⋮	⋮	⋮
x_{34}	0.454 9	0.450 1	0.018 9	0.011 6
x_{35}	0.498 2	0.473 3	0.009 6	0.010 2
x_{36}	0.712 6	0.719 3	0.018 3	0.011 7
x_{37}	0.321 1	0.312 0	0.014 4	0.012 0
x_{38}	0.503 6	0.502 0	0.022 3	0.020 1
y	0.590 3	0.588 9	0.024 4	0.021 0

表 5　NISEAE-VSG 模型性能说明

Table 5　Performance description of NISEAE-VSG Model

样本数	ARE	EIR_ARE
400	0.357 2	0
400＋40	0.278 1	0.221 4
400＋80	0.226 9	0.364 7
400＋160	0.192 7	0.460 5
400＋240	0.125 5	0.633 6
400＋320	0.317 5	0.111 9

表 6　加入不同的数量的虚拟样本比较不同方法的 ARE 和 R^2

Table 6　Comparing ARE and R^2 of different methods by adding different numbers of virtual samples

方法	虚拟样本数量									
	40		80		160		240		320	
	ARE	R^2	ARE	R^2	ARE	R^2	ARE	R^2	ARE	R^2
MTD	0.488 0	0.633 2	0.337 1	0.732 7	0.398 2	0.676 5	0.475 1	0.640 8	0.528 2	0.608 1
TTD	0.451 1	0.679 8	0.364 1	0.759 2	0.310 6	0.786 4	0.456 3	0.668 2	0.493 5	0.629 4
VAEVSG	0.376 2	0.758 4	0.336 4	0.795 3	0.291 1	0.836 1	0.264 2	0.885 4	0.384 3	0.738 6
GANVSG	0.391 5	0.732 5	0.318 8	0.809 4	0.258 7	0.871 8	0.290 8	0.815 5	0.405 4	0.728 6
NITAEVSG	0.353 9	0.771 8	0.312 7	0.814 6	0.250 3	0.869 0	0.296 5	0.821 8	0.336 8	0.789 5
NISEAEVSG	0.278 1	0.796 2	0.226 9	0.839 7	0.192 7	0.890 1	0.125 5	0.941 7	0.317 5	0.756 3

为了更好地表示预测输出与实际输出之间的分布，图 8 显示了 6 种 VSG 方法预测输出的散点图。在图 8 中，与其他 5 种 VSG 方法相比，所提方法数据点(蓝点)更接近于预期输出值(黑线)。说明所提方法在提高软测量模型精度方面是有效的。图 9 为 6 种 VSG 方法的输出预测误差箱型图。很明显，MTD 在这些 VSG 方法中具有最大的预测误差范围。相反，与其他模型相比，NISEAE-VSG 的误差区间要小得多。综上所述，所提方法显著提高了模型性能，验证了所提方法的有效性。

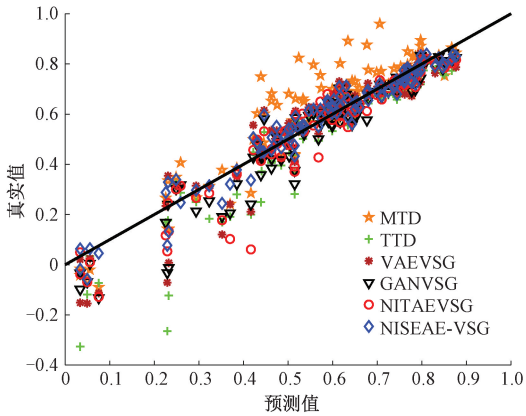

图 8　火力发电 6 种 VSG 方法测试输出的散点图

Fig. 8　Scatter plots of test outputs from 6 VSG methods for thermal power plant

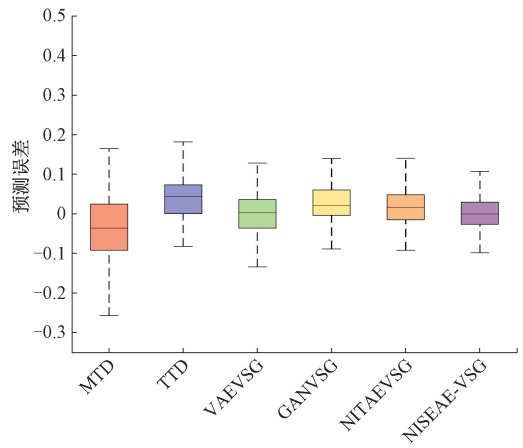

图 9　火力发电测试输出误差的箱形图

Fig. 9　Box plot of output error in thermal power testing

4　结　　论

本文提出的 NISEAE-VSG 模型能够处理复杂工业过程中的小样本问题。所提模型能够充分挖掘数据的隐藏特征，可以将输入和输出变量关联起来。为了验证所提方法的可行性和优越性，通过工业聚乙烯和火力发电的生产案例，验证了 NISEAE-VSG 对软测量模型性能的改善。仿真结果表明，与其他 6 种 VSG 方法相比，所提方法能够生成更加合理、可行的虚拟数据。通过 NISEAE-VSG 添加新生成的虚拟数据，提高软测量模型精度。由于单个样本是根据原始数据分布生成的，因此通常忽略了时空相关特征。未来的工作将考虑时空条件下的数据增强方法。

参考文献

[1]　张帅杰，郭小萍，臧春华，等. 多重注意机制及权重校正 LSTM 的 PVC 含水率预测[J]. 电子测量技术，2024，46(5)：83-90.
ZHANG SH J，GUO X P，ZANG CH H，et al. Prediction of PVC moisture content by multiple attention mechanism and weight correction LSTM[J]. Electronic Measurement Technology，2024，46(5)：83-90.

[2]　汤健，崔璨麟，夏恒，等. 面向复杂工业过程的虚拟样本生成综述[J]. 自动化学报，2024，50(1)：1-31.
TANG J，CUI C L，XIA H，et al. A survey of virtual sample generation for complex industrial processes[J]. Acta Automatica Sinica，2024，50(1)：1-31.

[3]　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[J]. Industrial & Engineering Chemistry Research，2020，59(40)：17977-17986.

[4]　陈忠圣，朱梅玉，贺彦林，等. 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[J]. 化工学报，2021，72(3)：1529-1538.
CHEN ZH SH，ZHU M Y，HE Y L，et al. Quantile regression CGAN based virtual samples generation and its applications to process modeling[J]. Journal of Chemical Industry and Engineering (China)，2021，72(3)：1529-1538.

[5]　ZHU Q X，LIU D P，XU Y，et al. Novel space projection interpolation based virtual sample generation for solving the small data problem in developing soft sensor[J]. Chemometrics and Intelligent Laboratory Systems，2021，217：104425.

[6]　FERNÁNDEZ A，GARCIA S，HERRERA F，et al. SMOTE for learning from imbalanced data：Progress and challenges，marking the 15-year anniversary[J]. Journal of Artificial Intelligence Research，2018，61：863-905.

[7]　LI D C，WU C S，TSAI T I，et al. Using mega-trend-diffusion and artificial samples in small data set learning for early flexible manufacturing system scheduling knowledge[J]. Computers & Operations Research，2007，34(4)：966-982.

[8]　LI D C，CHEN C C，CHANG C J，et al. A tree-based-trend-diffusion prediction procedure for small sample sets in the early stages of manufacturing

systems[J]. Expert Systems with Applications, 2012, 39(1): 1575-1581.

[9] HE Y L, WANG P J, ZHANG M Q, et al. 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[J]. Energy, 2018, 147: 418-427.

[10] 王丹丹，汤健，夏恒，等. 基于多目标 PSO 混合优化的虚拟样本生成[J]. 自动化学报，2024，50(4)：790-811.
WANG D D, TANG J, XIA H, et al. Virtual sample generation method based on hybrid optimization with multi-obiective PSO [J]. Acta Automatica Sinica, 2024, 50(4): 790-811.

[11] WU J Y, SHEN F F, YE L J. Data augmentation using time conditional variational autoencoder for soft sensor of industrial processes with limited data[J]. IEEE Transactions on Instrumentation and Measurement, 2024, 73: 1-14.

[12] XIE R M, JAN N M, HAO K, et al. Supervised variational autoencoders for soft sensor modeling with missing data [J]. IEEE Transactions on Industrial Informatics, 2019, 16(4): 2820-2828.

[13] KRISHNAN G S, KAMATH S. A novel GA-ELM model for patient-specific mortality prediction over large-scale lab event data [J]. Applied Soft Computing, 2019, 80: 525-533.

[14] LIU Y, GAO Z L, CHEN J H. Development of soft-sensors for online quality prediction of sequential-reactor-multi-grade industrial processes[J]. Chemical Engineering Science, 2013, 102: 602-612.

[15] ZHANG X M, HE B C, ZHU H. Y, et al. Information complementary fusion stacked autoencoders for soft sensor applications in multimode industrial processes [J]. IEEE Transactions on Industrial Informatics, 2023, 20(1): 106-116.

[16] YE T, XU Y, ZHU Q X, et al. Novel virtual sample generation using target-relevant autoencoder for small data-based soft sensor [J]. IEEE Transactions on Instrumentation and Measurement, 2021, 70: 1-10.

作者简介

郭小萍，博士，教授，主要研究方向为基于数据驱动的复杂过程故障诊断和软测量。

E-mail：gxp2001@sina.com

赵霄丰，硕士研究生，主要研究方向为基于数据驱动的复杂过程软测量。

E-mail：18640382503@163.com

李元(通信作者)，博士，教授，主要研究方向为基于数据驱动的复杂过程故障诊断。

E-mail：li-yuan@mail.tsinghua.edu.cn