DOI: 10.13382/j.jemi.B2205928

基于改进生成对抗网络的动液面建模数据扩充*

王 通 陈延彬
(沈阳工业大学电气工程学院 沈阳 110870)

摘 要:针对采用生成对抗网络进行油井生产参数数据生成时,部分生成数据不符合油井生产过程特性,导致动液面软测量建模质量不高的问题,提出一种基于专家诊断的生成对抗网络油井动液面软测量建模数据扩充方法。在判别器基于真实数据与生成数据得到原始损失值后,结合油井生产的机理过程对生成数据的合理性进行专家诊断,检测判别器判别结果。对错误结果进行补偿并加入生成器与判别器的损失函数中进行后续对抗训练,从而生成较优的符合油井生产过程特性的动液面软测量建模样本数据。通过仿真实验,生成数据补充到软测量建模的训练数据中能够提高动液面的预测精度,均方根误差降低了5.99%。表明加入专家诊断模块后生成器生成数据质量更高,能够更好地满足油田生产需求。

关键词: 数据扩充;动液面;生成对抗网络
中图分类号: TN919.5;TP183 **文献标识码**: A **国家标准学科分类代码**: 520.2060

Dynamic liquid level modeling data augmentation based on improved generative adversarial networks

Wang Tong Chen Yanbin
(School of Electrical Engineering, Shenyang University of Technology, Shenyang 110870, China)

Abstract: In using generative adversarial networks to generate oil well production parameter data, this method causes the inconsistency between partially generated data characteristics and characteristics of oil well production process, which leads to the low quality of soft sensor modeling of dynamic liquid level. This paper presents an expansion method of soft sensor modeling data of oil well dynamic liquid level based on expert diagnosis-wasserstein generative adversarial networks. After the discriminator obtains the original loss value based on the real data and generated data, the rationality of the generated data is diagnosed by the expert diagnosis module in combination with the mechanism process of oil well production, and the discriminator judgment results are detected. The error results are compensated and added to the loss functions of the generator and discriminator for subsequent confrontation training, thus the better soft sensor modeling sample data of dynamic liquid level which consistent with the characteristics of oil well production process is generated. Through simulation experiments, the prediction accuracy of the dynamic liquid level improved by adding the generated data to the training data of soft sensor modeling, and the root mean square error is reduced by 5.99%. It shows that the data generated by the generator after adding the expert diagnosis module has higher quality and can better meet the production needs of the oilfield.

Keywords: data augmentation; dynamic liquid level; generative adversarial networks

0 引 言

石油作为工业化进程的基础,直接影响着一国的经济、民生以及军事发展。我国的采油现场大多使用游梁式抽油机作为主要的生产工具,在生产过程中动液面会直接影响整个采油的效益,采用动液面来指导抽油机生产具有科学依据。目前动液面数据测量的主要方法为回声法,但回声法不能进行实时监测,需要人工定期到井口进行测量且操作繁琐,因此不能及时对油田生产情况进行分析。针对传统方法的不足,近些年许多专家学者提出以软测量的方式来解决上述问题[1-4]。软测量模型的

收稿日期: 2022-10-19 Received Date: 2022-10-19
*基金项目:国家自然科学基金(62173073)项目资助

质量与历史数据是否完备、信息量是否充足密切相关。但是油田现场生产过程属于不可重复过程,历史数据存在不平衡和缺失的情况,有些油井也存在数据稀少的问题。在此基础上建立的软测量模型往往达不到预期效果,因此扩充建模数据成为缓解上述问题的重要方法。

传统在对序列数据进行扩充的问题上,基本方法主要有几何变换、窗口裁剪、添加噪声等,这些方法简单直观方便,但未考虑到数据集整体的分布特点[5]。近几年生成对抗网络(generative adversarial networks, GAN)发展迅速,其以博弈的思维,生成式的模型在图像等领域都取得了不错的成绩[6-10]。由于 GAN 没有使用变分下界,如果判别器训练良好,那么生成器可以完美地学习到训练样本的概率分布[11]。基于这样的特点,GAN 在序列数据的生成上也逐渐得到应用,文献[12]针对电力系统拓扑改变后频率估计模型需要更新而训练样本不足的问题,使用 GAN 产生了大量相似样本。文献[13]针对海杂波数据稀少的问题,通过使用两个相同网络结构的 GAN 网络分别生成海杂波数据的实部和虚部,而后再合成得到了生成数据。文献[14]针对窃电检测的数据不平衡问题,通过基于 Wasserstein 距离的生成对抗网络(Wasserstein generative adversarial networks, WGAN)生成与真实窃电样本具有相近分布的合成样本。文献[15]针对新建光伏发电站原始数据匮乏导致光伏功率预测精度低的问题,使用带有梯度惩罚的 WGAN(Wasserstein generative adversarial networks-gradient penalty, WGAN-GP)实现训练集的增强,并提高了模型的预测精度。文献[16]针对油井现场历史数据不足,导致现有基于数据驱动的预测方法建模困难的问题,使用 GAN 对多尺度状态特征进行生成。然而上述文献在进行数据生成的时候,都是基于有限的历史数据依靠判别器的结果指导生成器进行学习。仅通过数据训练,会导致生成数据存在不符合实际工业过程的缺陷。进而使训练的软测量模型泛化性能变差。在复杂工业过程中,生产参数之间存在机理关系,生成的数据也应满足其所处工况的机理条件。因此,生成符合工业过程特性的数据,就必需结合相应复杂工业生产过程的先验知识。但现阶段中,基于生产规则来判断所生成的数据是否合理,进而引导 GAN 训练的研究还少有涉及。

为更好地建立动液面软测量模型,解决采用 WGAN 生成油井生产参数数据时,部分生成数据不符合油井生产过程特性的问题,提出一种基于专家诊断的生成对抗网络(expert diagnosis-wasserstein generative adversarial networks, ED-WGAN)油井动液面软测量建模数据扩充方法。根据油井生产的机理过程,设计油井生产参数专家诊断模块,利用该模块对生成数据是否合理进行检验,并得到判别器误判时的补偿损失,然后将其加入到对抗网络的损失函数中,以此引导训练过程,从而更有效地生成适合生产工况的数据。最后选取辽河油田某井历史数据对所提方法进行实验仿真验证。

1　生成对抗网络

1.1　GAN

GAN 是 Goodfellow[17] 在 2014 年提出的一种生成数据的深度学习模型,由两个神经网络组成:生成器(generator,G)和判别器(discriminator,D)。GAN 基本流程如图 1 所示。

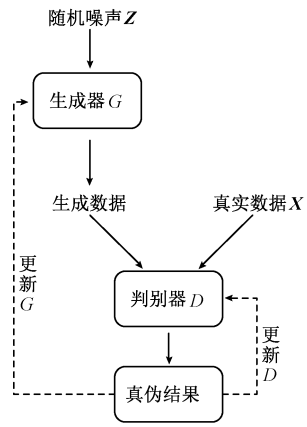

图 1　生成对抗网络流程

Fig. 1　Flow chart of generative adversarial networks

真实数据 $\boldsymbol{X}$ 来源于分布 p_r,随机噪声 $\boldsymbol{Z}$ 来源于先验分布 p_g,判别器与生成器可以被统一为一个目标函数,如式(1)所示。

$$\min_G \max_D V(D,G) = E_{X\sim P_r}[\lg D(\boldsymbol{X})] + E_{Z\sim P_g}[\lg(1 - D(G(\boldsymbol{Z})))] \tag{1}$$

式中:$E(\cdot)$ 为期望,$G(\boldsymbol{Z})$ 为生成数据,$D(G(\boldsymbol{Z}))$、$D(\boldsymbol{X})$ 分别为判别器对两种输入数据的判别输出。

式(1)分两步完成:先优化 D,目的是尽可能使判别器对输入数据做出正确判断;然后优化 G,目的是尽可能让生成器生成真实的数据,两者不断对抗训练,最终达到纳什均衡[18-19]。

1.2　WGAN

实际应用中,GAN 存在梯度消失,模型难以训练的问题[20]。WGAN[21]提出利用 Wasserstein 距离代替 JS 散度,作为度量分布差异的方法。Wasserstein 公式如式(2)所示:

$$W(p_r,p_g) = \inf_{\gamma\in\Pi(p_r,p_g)} E_{(x,y)\sim\gamma}[\,\|x - y\|\,] \tag{2}$$

式中:$\Pi(p_r,p_g)$ 是分布 p_r 和 p_g 的任意连接方式,$W(p_r,p_g)$ 是 $\gamma(x,y)$ 期望的下确界。

式(2)中 Wasserstein 距离中的下确界无法直接求

解,故采用其 Kantorovich-Rubinstein 对偶形式,如式(3)所示:

$$W(p_r,p_g)=\sup_{\|D\|_L\leqslant 1}E_{\boldsymbol{X}\sim P_r}[D(\boldsymbol{X})]-E_{\boldsymbol{Z}\sim P_g}[D(G(\boldsymbol{Z}))] \tag{3}$$

式中:$\|D\|_L\leqslant 1$ 表明判别器网络需要满足 1-Lipshitz 条件限制[22]。WGAN 通过参数裁剪的方式得以实现。

根据上述内容,WGAN 判别器与生成器的损失函数如式(4)、(5)所示。

$$L_D=E_{\boldsymbol{Z}\sim P_g}[D(G(\boldsymbol{Z}))]-E_{\boldsymbol{X}\sim P_r}[D(\boldsymbol{X})] \tag{4}$$

$$L_G=-E_{\boldsymbol{Z}\sim P_g}[D(G(\boldsymbol{Z}))] \tag{5}$$

2　基于 ED-WGAN 的油井生产参数扩充方法

WGAN 缓解了 GAN 容易出现的梯度消失与模式崩溃的问题,增加了多样性,使得生成数据在统计规律上基本符合真实数据。但所生成的数据,却存在不符合油井生产过程特性的问题。WGAN 所生成的部分数据如表 1 所示。

其中 1、2 组数据之间在套压数值发生极大变化的情况下,液量提高并且动液面高度也略微提高,但 40%的泵效降低到 25%,不合常理;3、4 组的套压数据出现了负数情况;5、6 组较符合,液量下降的同时泵效也相对变小,同时动液面高度略微上升。由于不合理的生成数据会影响建模质量,其占比越大软测量模型预测误差越大。因此,提升生成数据的整体质量能够有效提高软测量模型质量。

表 1　WGAN 所生成的部分数据

Table 1　Part of the data generated by WGAN

序号	液量/m³	油压/MPa	套压/MPa	电流比	泵效/%	动液面/m
1	9.127	0.327	0.645	1.274	40.145	870.21
2	10.296	0.439	2.120	1.866	25.040	904.34
3	14.621	0.062	−0.290	1.003	58.269	2 048.41
4	11.051	0.054	−0.239	0.887	43.649	1 956.48
5	10.639	0.442	1.017	1.477	34.527	1 473.59
6	8.700	0.431	0.904	1.277	31.202	1 500.30

本文根据油井生产的机理过程,提出一种基于 ED-WGAN 的油井动液面软测量建模数据扩充方法,其中,设计油井生产参数专家诊断模块诊断生成器所输出的生成数据。方法整体结构图如图 2 所示。

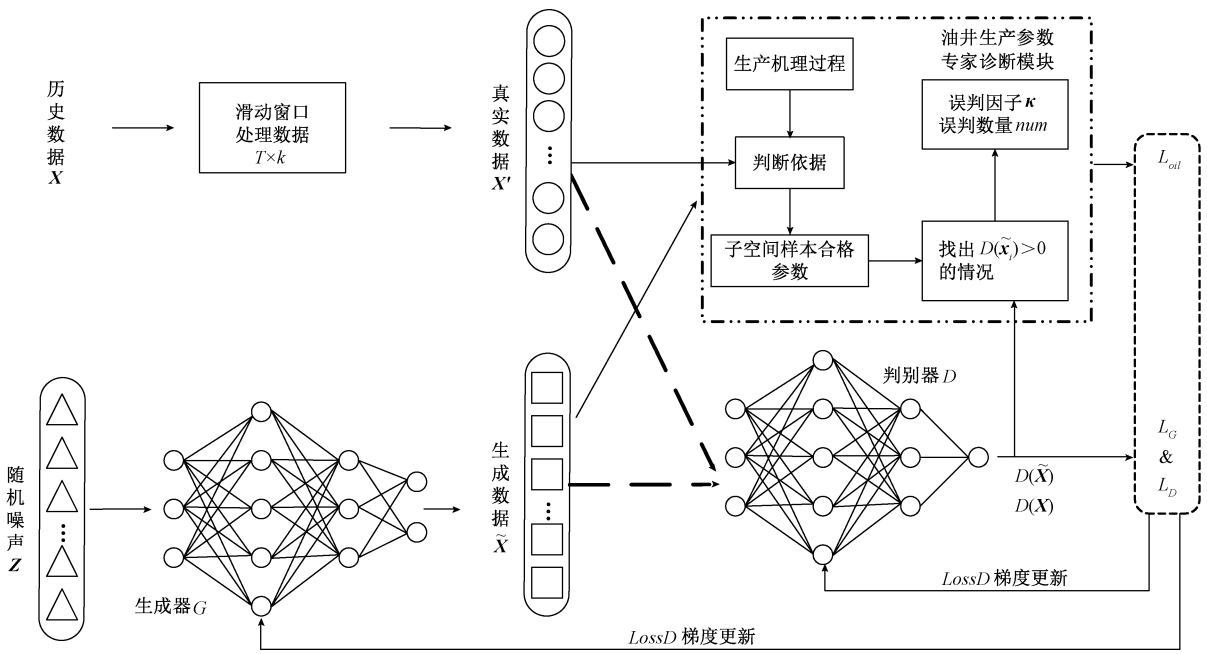

图 2　ED-WGAN 结构

Fig. 2　Structure of ED-WGAN

首先,将随机噪声 $\boldsymbol{Z}$ 输入生成器 G 得到生成数据 $\tilde{\boldsymbol{X}}=G(\boldsymbol{Z})$,判别器 D 基于真实数据 $\boldsymbol{X}$ 与生成数据 $\tilde{\boldsymbol{X}}$ 得到两组输出结果 $D(\boldsymbol{X})$ 和 $D(\tilde{\boldsymbol{X}})$,并最终计算得到 L_D 和 L_G。其次,专家诊断模块依据油井生产的机理过程,通过真实数据 $\boldsymbol{X}$ 来判断生成数据 $\tilde{\boldsymbol{X}}$ 是否合理,通过比较 $D(\boldsymbol{X})$ 和 $D(\tilde{\boldsymbol{X}})$ 的结果来确定判别器对不合理的生成数据是否判断正确,并得到补偿损失值。最后判别器损失函数加上补偿损失值,生成器损失函数减去补偿损失值,两者通过对抗训练,使生成器能够生成更符合油井生产过程特性的生成数据。

2.1 油井生产数据预处理

设油井生产数据 $\boldsymbol{X}=(\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_i,\cdots,\boldsymbol{x}_n)^{\mathrm{T}}\in\mathbb{R}^{n\times k}$，其中 $\boldsymbol{x}_i=(x_{i,1},x_{i,2},x_{i,3},\cdots,x_{i,k})$ 包含 k 个元素，为多维时间序列数据，每组数据之间具有连续性，$\boldsymbol{x}_i$ 内 k 个参数具有相关性，将历史数据按时间顺序单组输入作为真实数据训练后所生成的生成数据不具备时间连续性，无法体现油井参数之间的内在机理关系。对此，采用滑动窗口方法将油井生产数据按照时间顺序划分为多个生产数据子空间。

如图 3 所示，滑动窗口长度设为 T，k 为特征个数，窗口内 $T\times k$ 维的历史数据子空间构成一个输入到对抗网络的真实数据，其包含 $T\times k$ 时间段的连续油井生产参数，窗口滑动步长设为 1。经上述窗口滑动数据处理后，成为训练 ED-WGAN 的真实数据 $\boldsymbol{X}'=(\boldsymbol{x}'_1,\boldsymbol{x}'_2,\cdots,\boldsymbol{x}'_i,\cdots,\boldsymbol{x}'_{n-T+1})^{\mathrm{T}}\in\mathbb{R}^{(n-T+1)\times(T\times k)}$，其中任意子空间 $\boldsymbol{x}'_i=(\boldsymbol{x}_i,\boldsymbol{x}_{i+1},\cdots,\boldsymbol{x}_{i+T-1})^{\mathrm{T}}$ 包含 $T\times k$ 个元素。

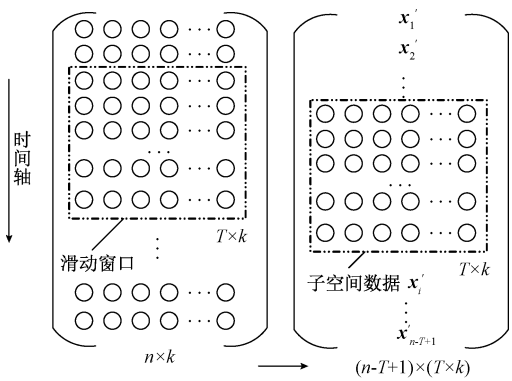

图 3 窗口滑动示意图

Fig. 3 Diagram of window sliding

2.2 油井生产参数专家诊断模块

WGAN 进行油井数据生成的过程中，部分生成数据不符合真实数据所处油井的生产过程特性，但判别器的输出结果却倾向于判断为真实数据，此类数据会影响软测量模型的训练效果。因此结合油井生产机理过程，设计一种油井生产参数专家诊断模块，对 WGAN 所生成的数据进行判别，使最终的训练结果往合理的方向发展。

1）油井生产机理过程

在油井生产的过程中，不同生产参数之间相互影响，相互制约。目前，许多专家和学者通过对油井生产过程的分析，得到一些生产参数之间的机理关系[23-24]，如下所示。

悬点静载荷可用式（6）、（7）来计算。

上行程：

$$W_{ju}=9.81q_{rl}L+A_pL_f\rho_L+10^6(p_t-p_c)A_p \tag{6}$$

下行程：

$$W_{jd}=9.81q_{rl}L \tag{7}$$

式中：W_{ju} 为上行程悬点静载荷，N；ρ_L 为油管内流体密度，kg/m^3；q_{rl} 为每米抽油杆在液体中的重力，kgf/m；L_f 为动液面高度，m；A_p 为柱塞面积，m^2；p_t 为油管压力，MPa；p_c 为套管压力，MPa；W_{jd} 为下行程悬点静载荷，N。

由式（6）、（7），可得到上下行程的载荷差：

$$\Delta W=A_pL_f\rho_L+10^6(p_t-p_c)A_p \tag{8}$$

最终得到动液面的计算公式：

$$L_f=\frac{\Delta W-10^6(p_t-p_c)A_p}{A_p\rho_L} \tag{9}$$

产液量与动液面和泵效的数学关系为：

$$Q=c(L_f-L_s) \tag{10}$$

$$\eta=Q/Q_c \tag{11}$$

式中：Q 为实际日产液量，m^3；L_s 为静液面高度，m；c 为常数；η 为泵效，无因次；Q_c 为理论日产液量，m^3。

2）专家诊断模块

假定每次迭代训练时，抽取 m 个 $T\times k$ 维高斯分布的噪声数据 $\boldsymbol{Z}$，通过生成器后生成 m 个子空间的生成数据 $G(\boldsymbol{Z})=\tilde{\boldsymbol{X}}=(\tilde{\boldsymbol{x}}_1,\tilde{\boldsymbol{x}}_2,\cdots,\tilde{\boldsymbol{x}}_i,\cdots,\tilde{\boldsymbol{x}}_m)^{\mathrm{T}}\in\mathbb{R}^{m\times(T\times k)}$，每一个子空间的生成数据 $\tilde{\boldsymbol{x}}_i=(\tilde{\boldsymbol{x}}_{i,1},\tilde{\boldsymbol{x}}_{i,2},\cdots,\tilde{\boldsymbol{x}}_{i,T})^{\mathrm{T}}$ 可表示为：

$$\tilde{\boldsymbol{x}}_i=\begin{pmatrix}\tilde{x}^{\,1}_{i,1} & \tilde{x}^{\,2}_{i,1} & \cdots & \tilde{x}^{\,k}_{i,1}\\ \tilde{x}^{\,1}_{i,2} & \tilde{x}^{\,2}_{i,2} & \cdots & \tilde{x}^{\,k}_{i,2}\\ \vdots & \vdots & \cdots & \vdots\\ \tilde{x}^{\,1}_{i,t} & \tilde{x}^{\,2}_{i,t} & \cdots & \tilde{x}^{\,k}_{i,t}\\ \vdots & \vdots & \cdots & \vdots\\ \tilde{x}^{\,1}_{i,T} & \tilde{x}^{\,2}_{i,T} & \cdots & \tilde{x}^{\,k}_{i,T}\end{pmatrix}_{(T\times k)}$$

其中，$i=1,2,3,\cdots,m$，表示生成数据的子空间数；$t=1,2,3,\cdots,T$，表示组数；k 为特征个数。

对式（10）和（11）分别改写成如下形式：

$$L_s=L_f-Q/c \tag{12}$$

$$Q_c=Q/\eta \tag{13}$$

基于真实数据 $\boldsymbol{X}$ 以及抽油井的生产参数 A_p、ρ_L 和 c，代入式（8）、（12）和（13）分别求出真实数据所处油井上下载荷差 ΔW、静液面高度 L_s 以及理论日产液量 Q_c 的变化范围[$\Delta W_{\min},\Delta W_{\max}$]、[$L_{s\min},L_{s\max}$]和[$Q_{c\min},Q_{c\max}$]。

将所求的变化范围与式（9）~（11）及 A_p、ρ_L 和 c 组成判断依据，生成数据通过判断依据可依次求得在相应套压与油压下的动液面、产液量与泵效的合理范围。

将生成数据 $\tilde{\boldsymbol{x}}_i$ 中第 t 组数据的油压 $\tilde{x}^{\,2}_{i,t}$ 与套压 $\tilde{x}^{\,3}_{i,t}$ 代入式（9），根据已求出的真实数据上下载荷差变化范围[$\Delta W_{\min},\Delta W_{\max}$]以及抽油井生产参数 A_p 和 ρ_L，得到

$\tilde{\boldsymbol{x}}_i$ 的第 t 组动液面高度合理范围 $[L_{f\min}^{i,t}, L_{f\max}^{i,t}]$。把 $[L_{f\min}^{i,t}, L_{f\max}^{i,t}]$ 与已知的静液面高度变化范围 $[L_{s-\min}, L_{s-\max}]$ 和 c 代入式(10),可得实际产液量的合理范围 $[Q_{\min}^{i,t}, Q_{\max}^{i,t}]$。同样把 $[Q_{\min}^{i,t}, Q_{\max}^{i,t}]$ 与已知的理论产液量变化范围 $[Q_{c-\min}, Q_{c-\max}]$ 代入式(11),可得泵效的合理范围 $[\eta_{\min}^{i,t}, \eta_{\max}^{i,t}]$。

根据已求得的合理范围,对第 t 组生成数据的动液面高度 $\tilde{x}_{i,t}^{6}$、产液量 $\tilde{x}_{i,t}^{1}$ 以及泵效 $\tilde{x}_{i,t}^{5}$ 进行合理性判定,设计 3 个参数判定因子,用于保存生成数据的判定结果,参数定义如下。

$$\boldsymbol{A}(i,t)=\begin{cases}0, & \tilde{x}_{i,t}^{6}\notin[L_{f\min}^{i,t}, L_{f\max}^{i,t}]\\ 1, & \tilde{x}_{i,t}^{6}\in[L_{f\min}^{i,t}, L_{f\max}^{i,t}]\end{cases}\tag{14}$$

$$\boldsymbol{B}(i,t)=\begin{cases}0, & \tilde{x}_{i,t}^{1}\notin[Q_{\min}^{i,t}, Q_{\max}^{i,t}]\\ 1, & \tilde{x}_{i,t}^{1}\in[Q_{\min}^{i,t}, Q_{\max}^{i,t}]\end{cases}\tag{15}$$

$$\boldsymbol{C}(i,t)=\begin{cases}0, & \tilde{x}_{i,t}^{5}\notin[\eta_{\min}^{i,t}, \eta_{\max}^{i,t}]\\ 1, & \tilde{x}_{i,t}^{5}\in[\eta_{\min}^{i,t}, \eta_{\max}^{i,t}]\end{cases}\tag{16}$$

其中,$\boldsymbol{A}$ 参数保存动液面的判定结果,$\boldsymbol{B}$ 参数保存产液量的判定结果,$\boldsymbol{C}$ 参数保存泵效的判定结果。按式(14)~(16)判断生成数据的动液面、产液量和泵效数值是否超限,并保存。

为了判断生成数据子空间 $\tilde{\boldsymbol{x}}_i$ 样本是否符合生产过程特性,设计子空间样本合格参数 $\boldsymbol{K}$,参数定义如下。

$$\boldsymbol{K}(i)=\prod_{t=1}^{T}[\boldsymbol{A}(i,t)\cdot\boldsymbol{B}(i,t)\cdot\boldsymbol{C}(i,t)]\tag{17}$$

其中,T 为子空间长度,如子空间内样本存在不符合生产特性的样本,则该子空间样本合格参数 $\boldsymbol{K}(i)=0$。

定义误判因子 $\boldsymbol{\kappa}$,如式(18)所示:

$$\boldsymbol{\kappa}(i)=\begin{cases}1, & (D(\tilde{\boldsymbol{x}}_i)>0)\ \&(\boldsymbol{K}(i)==0)\\ 0, & \text{其他}\end{cases}\tag{18}$$

由式(17)可知,$\boldsymbol{K}(i)=0$ 为子空间中存在不符合油井生产过程特性的生成数据,此时如 $D(\tilde{\boldsymbol{x}}_i)>0$ 时,误判因子 $\boldsymbol{\kappa}$ 反应判别器结果与专家诊断模块是否偏离。

参考 WGAN 损失函数的形式,用误判程度和的均值定义补偿损失。

$$L_{oil}=\frac{\sum_{i=1}^{m}[\boldsymbol{\kappa}(i)\cdot D(\tilde{\boldsymbol{x}}_i)]}{num+\zeta}\tag{19}$$

其中,$num=sum(\boldsymbol{\kappa})$ 为 m 个子空间的生成数据中误判的数量,ζ 是一个很小的数,为了防止分母为 0。按式(19),可得最终的补偿损失,将其作为判别器对生成数据进行判决时的补偿。整体的流程图如图 4 所示。

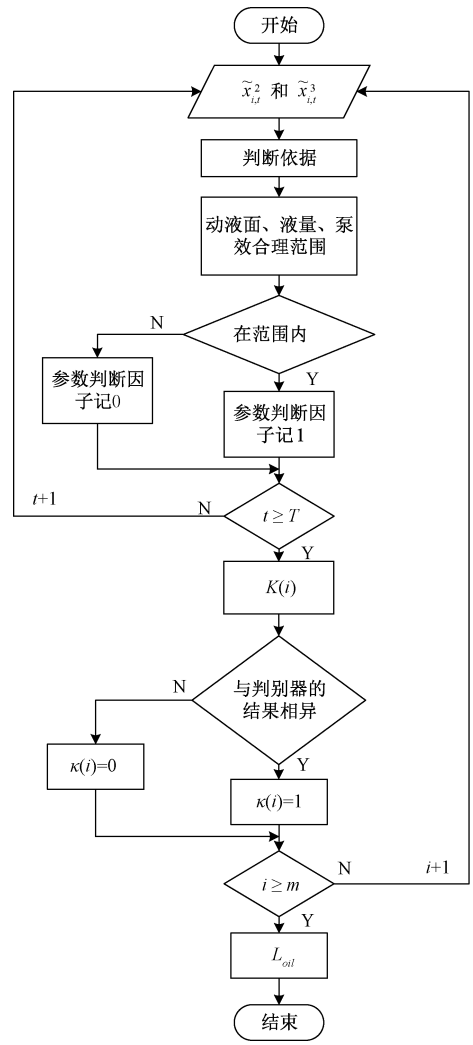

图 4　专家诊断模块流程

Fig. 4　Flow chart of expert diagnosis module

2.3　损失函数的设计

在 WGAN 中,生成器与判别器的损失函数可被统一为如式(20)所示。

$$\min_{G}\max_{D}V(D,G)=E_{\boldsymbol{X}\sim P_r}[D(\boldsymbol{X})]-E_{\boldsymbol{Z}\sim P_g}[D(G(\boldsymbol{Z}))]\tag{20}$$

判别器的目的是通过最大化 $V(D,G)$,以近似找到真实数据与生成数据之间的 Wasserstein 距离。为了得到最大化的 $V(D,G)$,可以让 $E_{\boldsymbol{X}\sim P_r}[D(\boldsymbol{X})]$ 梯度上升,让 $E_{\boldsymbol{Z}\sim P_g}[D(G(\boldsymbol{Z}))]$ 梯度下降。采用梯度下降法来训练时,判别器的损失函数由式(4)所示。

生成器最小化 $V(D,G)$,由于 $E_{\boldsymbol{X}\sim P_r}[D(\boldsymbol{X})]$ 与生成器无关,因此为让 $E_{\boldsymbol{Z}\sim P_g}[D(G(\boldsymbol{Z}))]$ 梯度上升,采用梯度下降法来训练时,生成器的损失函数可由式(5)所示。

油井生产参数之间存在着一定的机理关系,在训练

的过程中可以通过机理过程来判断判别器的输出是否合理。补偿损失函数 L_{oil} 能够反映判别器误判程度,其值越大表示专家诊断模块发现判别器对生成数据进行判决时出现的偏差越大。在训练过程中,可沿着使 L_{oil} 尽可能小的方向进行梯度优化。式(20)表示的目标函数可修改为:

$$\min_{oil}\min_{G}\max_{D}V(D,G,oil) = E_{\boldsymbol{X}\sim P_r}[D(\boldsymbol{X})] - E_{\boldsymbol{Z}\sim P_g}[D(G(\boldsymbol{Z}))] - L_{oil} \tag{21}$$

此时 ED-WGAN 判别器是最大化 $V(D,G,oil)$,采用梯度下降法来训练时,损失函数可表示为:

$$LossD = -E_{\boldsymbol{X}\sim P_r}[D(\boldsymbol{X})] + E_{\boldsymbol{Z}\sim P_g}[D(G(\boldsymbol{Z}))] + L_{oil} \tag{22}$$

ED-WGAN 生成器是最小化 $V(D,G,oil)$,采用梯度下降法来训练时,损失函数可表示为:

$$LossG = -E_{\boldsymbol{Z}\sim P_g}[D(G(\boldsymbol{Z}))] - L_{oil} \tag{23}$$

2.4　生成器与判别器的结构设计

在生成对抗网络中,时序数据采用卷积神经网络结构具有诸多不便[25],因此采用全连接神经网络作为生成器和判别器的结构。

1)生成器网络由 4 层神经网络组成,包括输入层、隐含层和输出层。输入层数据为随机产生的高斯分布噪声数据 $\boldsymbol{Z}$,中间为两层隐含层,节点数分别为 120 和 240,为解决 ReLU 激活函数进入负区间后,导致神经元不学习的问题[26]。隐含层的激活函数选为带泄露修正线性单元(Leaky-ReLU)函数:

$$f(x) = \begin{cases} x & ,x > 0 \\ \lambda x & ,x \leqslant 0 \end{cases} \tag{24}$$

其中,λ 是一个极小的数,使得该函数的输出对负值输入有很小的坡度。

两个非线性激活函数如图 5 所示。

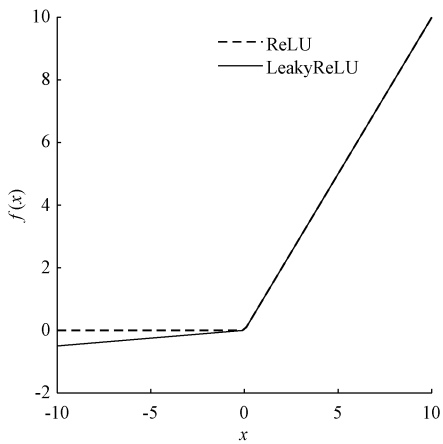

图 5　两种激活函数的数学形式

Fig. 5　The mathematical forms of two activation functions

输出层节点数为 36,无激活函数。

2)判别器网络同样为 4 层神经网络组成,包括输入层、隐含层和输出层。输入层接收生成数据和真实数据,为缓解梯度消失和梯度爆炸,确保多层神经网络的有效性[27],真实数据在输入判别器前,进行 0 均值标准化的操作:

$$x_{normalization} = \frac{x-\mu}{\sigma} \tag{25}$$

其中,μ 为均值,σ 为方差。

两层隐含层节点数分别为 120 和 240,激活函数为 Leaky-ReLU 函数,输出层节点数为 1,无激活函数。

在 WGAN 中,使用自适应矩估计算法(adaptive moment estimation,Adam)会导致梯度更新的不稳定,因此选择前向均方根梯度下降算法(root mean square prop,RMSprop)作为模型的优化器,按下式进行网络参数的优化:

$$\begin{cases} s_{dw} = \beta s_{dw} + (1-\beta)dW^2 \\ s_{db} = \beta s_{db} + (1-\beta)db^2 \\ W = W - \alpha \dfrac{dW}{\sqrt{s_{dw}} + \varepsilon} \\ b = b - \alpha \dfrac{db}{\sqrt{s_{db}} + \varepsilon} \end{cases} \tag{26}$$

其中,β 为梯度累积的一个指数,设置为 0.99;α 为学习速率,生成器设置为:0.000 5,判别器设置为:0.000 3;ε 是一个很小的数,为防止分母为 0,一般取值为 1×10^{-8}。

3　实验与分析

仿真实验平台建立在 Windows 系统 Python 软件下,电脑配置为 Intel Corei7-10750 H 2.6 GHz 处理器,16 G 内存,1 T 固态硬盘。

3.1　数据来源

选取辽河油田某井的 161 组历史生产参数数据,分别为:产液量、油压、套压、电流比、泵效和动液面。间隔抽取了其中 107 组为真实数据,经长度为 6 的滑动窗口转换为训练模型的 102 个真实子空间数据,子空间规模为 6×6。

3.2　实验结果分析

为验证方法的有效性,将 WGAN 与 ED-WGAN 进行对比,其中两种算法的随机数种子均设置为 2。两个模型均迭代 5 000 次,判别器先更新 5 次后生成器再更新一次。裁剪系数设置为 0.01。

通过仿真可以看出 Wasserstein 距离在改进前后的对比情况如图 6 所示。

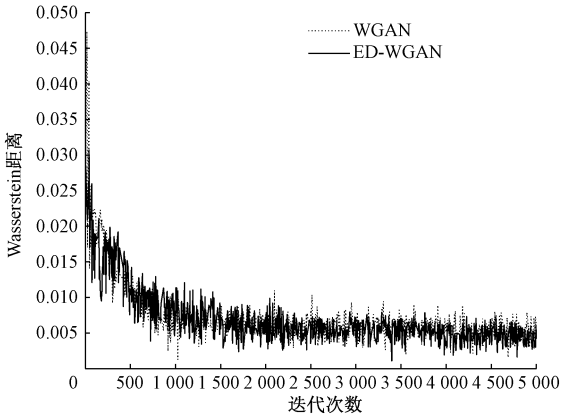

图 6　Wasserstein 距离对比

Fig. 6　Comparison chart of Wasserstein distance

由图 6 可看出,在 0~3 000 次的迭代中,改进前后模型的 Wasserstein 距离整体持续下降,模型处在学习阶段。在 3 000~5 000 次的迭代过程中,WGAN 整体波动稳定并维持在一定范围内不再下降,ED-WGAN 仍有较小的下降趋势,最后的 Wasserstein 距离也更低,说明 ED-WGAN 所生成数据与真实数据之间分布更接近。

记录 WGAN 与 ED-WGAN 模型每一轮迭代中所生成的数据。分别对比两者在第 5 次、3 000 次和 5 000 次迭代所生成的部分生成数据,如图 7~9 所示,图中虚线为训练网络的真实数据上下限。

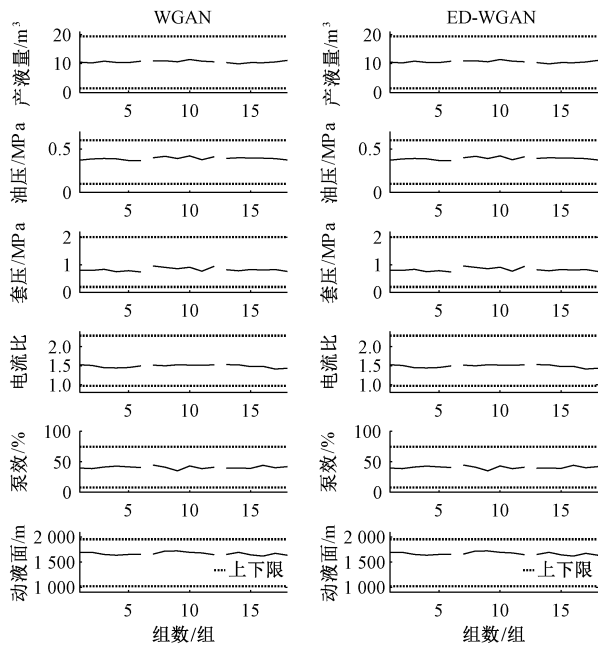

图 7　第 5 次迭代生成的部分数据

Fig. 7　Part of the data generated: Epoch = 5

由图 7 可看出,在训练初期第 5 次迭代时,WGAN 与 ED-WGAN 生成数据平稳,不存在明显不合理的情况,同时都缺乏多样性,此类数据对软测量的建模毫无作用。

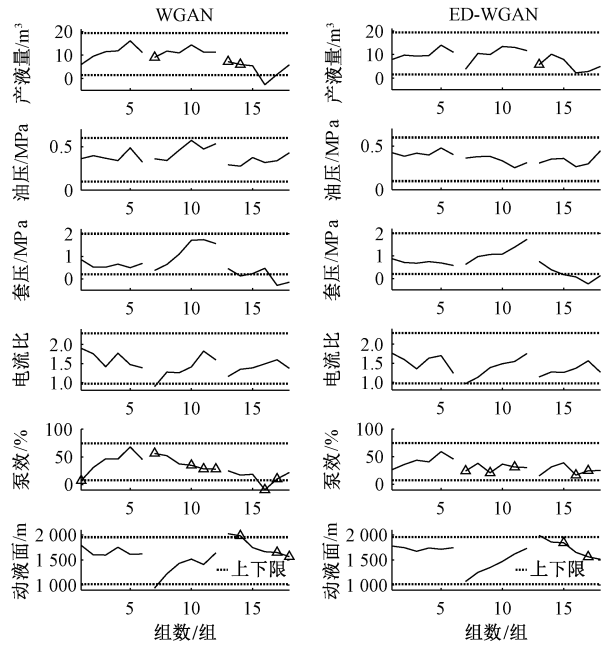

图 8　第 3 000 次迭代生成的部分数据

Fig. 8　Part of the data generated: Epoch = 3 000

随着迭代的进行,在训练中期生成数据的多样性得到了改善,但 WGAN 所生成的数据出现了较多不合理之处(Δ 标注),图 8 中 WGAN 生成数据的产液量、套压和泵效均出现了负值,该值在油井生产过程中不可能出现。此外,WGAN 生成数据组 1 的油压为 0.36 MPa,套压为 0.85 MPa,根据式(6)~(8)的机理公式可知,在理论产液量为[19.15 m^3, 39.00 m^3]、上下行程载荷差为[730.3 N,19 856.8 N]以及静液面高度为[1 006.4 m, 1 950.9 m]的条件下,计算得泵效数值的基本范围为[15.36%,31.28%],此时生成数据的泵效为 7.14%并不在区间内,可判断该数据不合理。图 8 中 ED-WGAN 生成数据仅套压存在负值,不合理情况也较少。

由图 9 可知,在第 5 000 次迭代时 WGAN 生成数据的套压出现了负值,泵效出现了超过 100%,且不符合油井生产过程特性的数据依旧存在。而图 9 中 ED-WGAN 生成数据都合乎要求。通过上述对比可以直观的看出,两种算法所得到数据的大致走势和形态相差不大,然而 ED-WGAN 所生成数据的合理性却比 WGAN 的更高。

为验证本文方法生成数据在动液面软测量建模过程中的有效性,将 107 组历史真实数据分别与 WGAN 和 ED-WGAN 各自生成的 20 个子空间即 120 组生成数据组成训练数据集 1 和训练数据集 2,将未采用数据扩充的 107 组历史数据单独组成训练数据集 3,基于上述 3 个训练数据集分别进行动液面软测量建模,测试数据集均采用另外 54 组历史真实数据。软测量模型采用 PSO-SVR

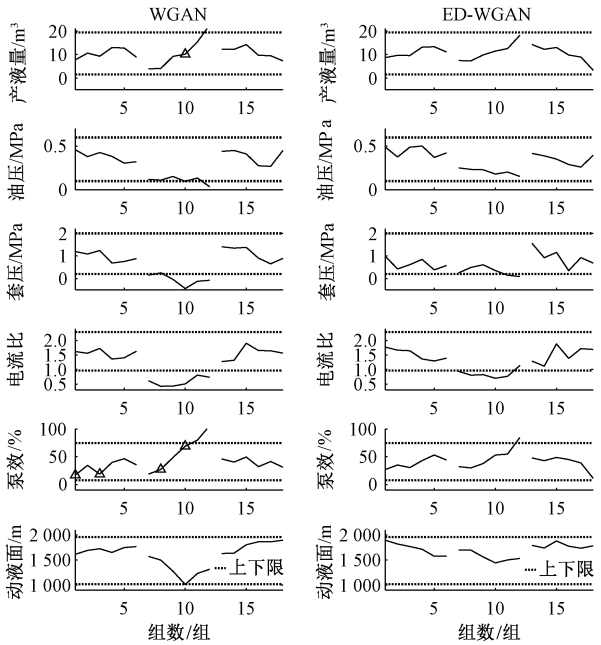

图9　第5 000次迭代生成的部分数据

Fig. 9　Part of the data generated: Epoch = 5 000

建立,其中PSO的参数设置如表2所示,以误差为适应度函数,搜索得到SVR的惩罚因子 $C=1\,568.7$,核函数因子 $g=0.976$。

表2　PSO参数设置表

Table 2　PSO parameter setting table

参数	数值
迭代次数	100
种群大小	40
C 的搜索区间	[1,2 000]
g 的搜索区间	[0.01,50]
学习常数 C_1	2
学习常数 C_2	1.7
权重因子	0.8

软测量模型性能评估指标选择MAE和RMSE:

$$MAE=\frac{1}{n}\sum_{i=1}^{n}|y_i-\hat{y}_i| \tag{27}$$

$$RMSE=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2} \tag{28}$$

结果对比结果如图10和表3所示。

可以看出,采用WGAN进行训练数据的扩充后,训练样本的数量得到了增加,但PSO-SVR的预测效果并未得到提升,反而由于数据质量较低,导致了模型能力下降,效果并不理想。采用ED-WGAN进行扩充后,在满足建模对数据量需求的基础上,PSO-SVR的预测准确度得到了提升,数据质量更优。

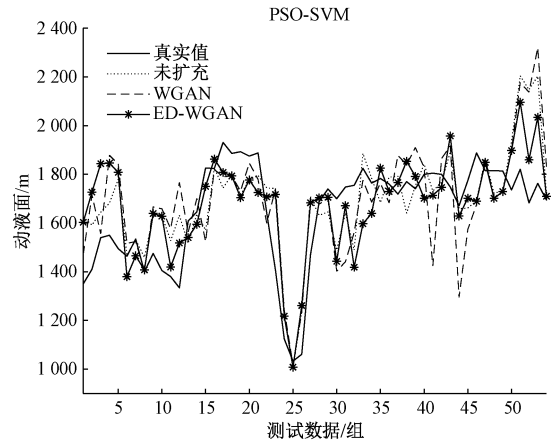

图10　软测量结果对比

Fig. 10　Comparison chart of soft sensor results

表3　误差结果对比表

Table 3　Comparison table of error result

	MAE	RMSE
未扩充	136.38	175.57
WGAN	165.47	213.32
ED-WGAN	135.02	165.65

为避免生成数据数量对建模结果产生干扰,测试在不同生成数据组数情况下,采用上述实验执行后的效果,对比两种模型所扩充的训练数据量对软测量模型的影响,如图11所示。

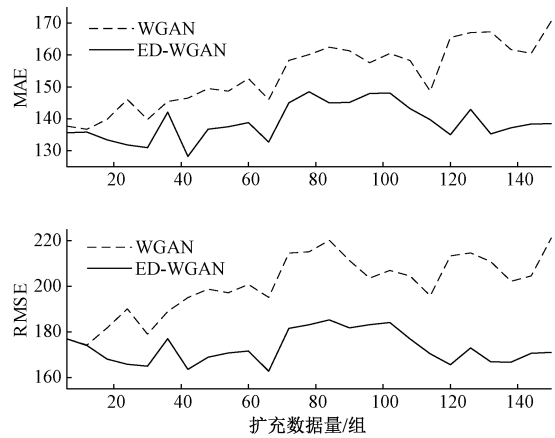

图11　不同扩充数据量下的评价指标对比

Fig. 11　Comparison chart of evaluation indexes under different expansion quantities

将不同数量生成数据加入原训练数据后建模并预测得出的MAE与RMSE指标对比图中可以看出,采用WGAN进行数据扩充后的模型质量均不如采用ED-WGAN进行相同数据扩充后的模型。由于训练数据的质量决定软测量模型的性能,当生成样本中优质数据数量

较高时,软测量模型性能会得到提升。通过上述实验,能够得出 ED-WGAN 生成样本数据质量优于 WGAN 生成样本且较为稳定,说明专家诊断模块起到作用,ED-WGAN 所生成的数据比 WGAN 更符合油井生产过程特性。本文为提高生成数据质量,在 WGAN 基础上添加专家诊断模块。因此,改进后算法训练时长较原算法略有增加,其中 WGAN 训练时长 4.108 h,ED-WGAN 训练时长 4.149 h。

4　结　论

为解决 WGAN 生成的部分油井生产参数数据不符合油井生产过程特性的问题,提出了一种基于 ED-WGAN 的油井动液面软测量建模数据扩充方法。构建专家诊断模块对生成数据从机理上进行合理性判决,使得生成的数据更符合生产过程特性,从而提高整体生成数据的质量。通过仿真结果的对比分析可以看出,专家诊断模块能够对判别器的结果进行判断,所得到的补偿损失也能对生成对抗网络的训练起到积极引导作用。最后采用 ED-WGAN 进行数据扩充后的 PSO-SVR 模型各项评估指标都更好,说明本文方法所生成的数据比 WGAN 质量更优,更符合油井生产过程特性。

参考文献

[1] 王通,段泽文,李琨. 基于改进 AdaBoost 的油井动液面自适应集成建模[J]. 电子测量与仪器学报,2017,31(8):1342-1348.
WANG T, DUAN Z W, LI K. Adaptive ensemble modeling for dynamic liquid level of oil well based on improved AdaBoost method [J]. Journal of Electronic Measurement and Instrumentation, 2017, 31 (8): 1342-1348.

[2] 李翔宇,高宪文,李琨,等. 基于多源信息特征融合的抽油井动液面集成软测量建模[J]. 化工学报,2016,67 (6):2469-2479.
LI X Y, GAO X W, LI K, et al. Ensemble soft sensor modeling for dynamic liquid level of oil well based on multi-source information feature fusion [J]. CIESC Journal, 2016, 67(6): 2469-2479.

[3] 王通,高宪文,刘文芳. 基于改进即时学习算法的动液面软测量建模[J]. 东北大学学报(自然科学版),2015,36(7):918-922.
WANG T, GAO X W, LIU W F. Soft sensor for determination of dynamic fluid levels based on enhanced just-in-time learning algorithm [J]. Journal of Northeastern University Natural Science, 2015, 36(7): 918-922.

[4] 王通,段泽文. 基于模糊评估自适应更新的油井动液面软测量建模[J]. 化工学报, 2019, 70 (12): 4760-4769.
WANG T, DUAN Z W. Soft sensor modeling for dynamic liquid level of oil well based on fuzzy inference adaptive updating [J]. CIESC Journal, 2019, 70 (12): 4760-4769.

[5] 葛轶洲,许翔,杨锁荣,等. 序列数据的数据增强方法综述[J]. 计算机科学与探索, 2021, 15 (7): 1207-1219.
GE Y ZH, XU X, YANG S R, et al. Survey on sequence data augmentation [J]. Journal of Frontiers of Computer Science and Technology, 2021, 15 (7): 1207-1219.

[6] SUN L Y, CHEN P, XIANG W, et al. SmartPaint:一种基于生成式对抗神经网络的人机协同绘画系统(英文)[J]. Frontiers of Information Technology & Electronic Engineering,2019,20(12):1644-1657.

[7] 刘雨溪,张铂,王斌. 基于生成式对抗网络的遥感图像半监督语义分割[J]. 红外与毫米波学报,2020,39(4):473-482.
LIU Y X, ZHANG B, WANG B. Semi-supervised semantic segmentation based on generative adversarial networks for remote sensing images [J]. Journal of Infrared and Millimeter Waves, 2020, 39(4): 473-482.

[8] 于希明,洪硕,于金祥,等. 可见光遥感图像船舶目标数据增强方法研究[J]. 仪器仪表学报,2020,41(11):261-269.
YU X M, HONG SH, YU J X, et al. Research on a ship target data augmentation method of visible remote sensing image [J]. Chinese Journal of Scientific Instrument, 2020,41(11):261-269.

[9] 王桂棠,林桢哲,符秦沈,等. 联合生成对抗网络的肺结节良恶性分类模型[J]. 仪器仪表学报,2020,41(11):188-197.
WANG G T, LIN ZH ZH, FU Q SH, et al. Joint generative adversarial network model for classification of benign and malignant pulmonary nodules [J]. Chinese Journal of Scientific Instrument, 2020,41(11):188-197.

[10] 王昕钰,王倩,程敦诚,等. 基于三级级联架构的接触网定位管开口销缺陷检测[J]. 仪器仪表学报, 2019, 40(10):74-83.
WANG X Y, WANG Q, CHENG D CH, et al. Detection of split pins defect in catenary positioning tube based on three-level cascade architecture [J]. Chinese Journal of Scientific Instrument, 2019,40(10):74-83.

[11] 刘建伟,谢浩杰,罗雄麟. 生成对抗网络在各领域应用研究进展[J]. 自动化学报,2020,46(12):2500-2536.
LIU J W, XIE H J, LUO X L. Research progress on

application of generative adversarial networks in various fields[J]. Acta Automatica Sinica, 2020,46(12):2500-2536.

[12] 赵冬梅,郑亚锐,谢家康,等. 基于轻量级梯度提升机和生成对抗网络的含风电电力系统频率稳定评估[J]. 电网技术,2022,46(8):3181-3193.
ZHAO D M, ZHENG Y R, XIE J K, et al. Frequency stability evaluation of power system containing wind power based on light gradient boosting machine and generative adversarial network [J]. Power System Technology, 2022, 46(8):3181-3193.

[13] 丁斌,夏雪,梁雪峰. 基于深度生成对抗网络的海杂波数据增强方法[J]. 电子与信息学报, 2021,43(7):1985-1991.
DING B, XIA X, LIANG X F. Sea clutter data augmentation method based on deep generative adversarial network [J]. Journal of Electronics & Information Technology, 2021, 43(7):1985-1991.

[14] 王德文,杨凯华. 基于生成式对抗网络的窃电检测数据生成方法[J]. 电网技术,2020,44(2):775-782.
WANG D W, YANG K H. A data generation method for electricity theft detection using generative adversarial network[J]. Power System Technology, 2020, 44(2): 775-782.

[15] 殷豪,张铮,丁伟锋,等. 基于生成对抗网络和LSTM-CSO的少样本光伏功率短期预测[J]. 高电压技术,2022,48(11):4342-4351.
YIN H, ZHANG ZH, DING W F, et al. Short term prediction of small sample photovoltaic power based on generative adversarial network and LSTM-CSO[J]. High Voltage Engineering,2022,48(11):4342-4351.

[16] 侯延彬,高宪文,李翔宇. 采油过程多尺度状态特征生成的有杆泵动态液面预测[J]. 化工学报,2019,70(S2):311-321.
HOU Y B, GAO X W, LI X Y. Prediction for dynamic liquid level of sucker rod pumping using generation of multi-scale state characteristics in oil field production[J]. CIESC Journal, 2019, 70(S2): 311-321.

[17] GOODFELLOW I J, POUGET-ABADIE J, MIRZA M, et al. Generative adversarial nets[C]. Proceedings of the 27th International Conference on Neural Information Processing Systems. Montreal, Canada: MIT Press, 2014:2672-2680.

[18] 罗佳,黄晋英. 生成式对抗网络研究综述[J]. 仪器仪表学报,2019,40(3):74-84.
LUO J, HUANG J Y. Generative adversarial network: An overview [J]. Chinese Journal of Scientific Instrument,2019, 40(3):74-84.

[19] 张鑫,缪楠,高继勇,等. 基于电子舌和WGAN-CNN模型的小麦贮存年限快速检测[J]. 电子测量与仪器学报,2021,35(6):176-183.
ZHANG X, MIAO N, GAO J Y, et al. Rapid detection of wheat storage year based on electronic tongue and WGAN-CNN model [J]. Journal of Electronic Measurement and Instrumentation, 2021, 35 (6): 176-183.

[20] ARJOVSKY M, BOTTOU L. Towards principled methods for training generative adversarial networks[J]. arXiv preprint arXiv:1701.04862, 2017.

[21] ARJOVSKY M, CHINTALA S, BOTTOU L. Wasserstein GAN[J]. Machine Learning, 2017, DOI: 10.48550/arXiv.1701.07875.

[22] 刘云鹏,许自强,和家慧,等. 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[J]. 电网技术, 2020,44(4):1505-1513.
LIU Y P, XU Z Q, HE J H, et al. Data augmentation method for power transformer fault diagnosis based on conditional Wasserstein generative adversarial network[J]. Power System Technology, 2020,44(4):1505-1513.

[23] 刘显赫. 有杆抽油机的功图数据采集与功图计量研究[D]. 沈阳:沈阳工业大学,2021.
LIU X H. Power diagram data acquisition and power diagram measurement of rod pumping unit [D]. Shenyang:Shenyang University of Technology,2021.

[24] 王通,高宪文,刘文芳. 自适应软测量方法在动液面预测中的研究与应用[J]. 化工学报,2014,65(12):4898-4904.
WANG T, GAO X W, LIU W F. Adaptive soft sensor method and application in determination of dynamic fluid levels[J]. CIESC Journal, 2014, 65(12): 4898-4904.

[25] 金晓航,许壮伟,孙毅,等. 基于生成对抗网络的风电机组在线状态监测[J]. 仪器仪表学报,2020,41(4):68-76.
JIN X H, XU ZH W, SUN Y, et al. Online condition monitoring of wind turbine based on generative adversarial network[J]. Chinese Journal of Scientific Instrument, 2020,41(4):68-76.

[26] XU J, LI Z, DU B, et al. Reluplex made more practical: Leaky ReLU[C]. 2020 IEEE Symposium on Computers and Communications (ISCC). IEEE, 2020: 1-7.

[27] 田娟秀,刘国才,谷珊珊,等. 医学图像分析深度学习方法研究与挑战[J]. 自动化学报,2018,44(3):401-424.
TIAN J X, LIU G C, GU SH SH, et al. Deep learning in medical image analysis and its challenges [J]. Acta

Automatica Sinica, 2018, 44(3): 401-424.

作者简介

王通, 1999 年于沈阳航空工业学院获得学士学位, 2004 年于大连理工大学获得硕士学位, 2015 年于东北大学获得博士学位, 现为沈阳工业大学副教授, 主要研究方向为复杂工业环境下的故障诊断和监控关键技术研究与应用。

E-mail: tykj_wt@126.com

Wang Tong received his B. Sc. degree from Shenyang Institute of Aeronautical Engineering in 1999, M. Sc. degree from Dalian University of Technology in 2004, and Ph. D. degree from Northeastern University in 2015, respectively. Now he is an associate professor in Shenyang University of Technology. His main research interests include the key technologies of fault diagnosis and monitoring in complexity industrial condition.

陈延彬(通信作者), 2016 年于南京理工大学泰州科技学院获得学士学位, 现为沈阳工业大学在读硕士研究生, 主要研究方向为软测量。

E-mail: c_ybin@126.com

Chen Yanbin (Corresponding author) received his B. Sc. degree from Taizhou Institute of Sci. &Tech. in 2016. Now he is a M. Sc. candidate in Shenyang University of Technology. His main research interest includes soft sensor.