基于 GAN 的小样本腐蚀失厚率数据增强方法

周俊炎，王竟成，杨小奎，舒畅，王津梅，张宸

（西南技术工程研究所，重庆 400039）

摘要：目的 对小样本腐蚀失厚率数据进行数据增强，实现数据扩充，以提升后续分析模型的预测精度，减轻过拟合程度，并提升模型的泛化能力。**方法** 利用生成对抗网络（Generative Adversarial Networks，GAN）扩充腐蚀失厚率数据，使数据分布更加全面。对生成数据进行降维可视化分析，探究生成数据与原始数据样本的分布规律，分析数据增强合理性，并从多个算法模型、多个评价指标角度对分析预测能力、泛化能力进行评估。**结果** 生成数据填补了原始数据在样本空间分布的薄弱环节，加入生成数据后，各机器学习算法模型得出的 MSE 均值为未加入生成数据的 61.72%~91.74%，皮尔逊均值为 99.01%~113.64%，预测准确度提升，结果关联性更强，模型泛化能力增强。**结论** GAN 能有效对小样本腐蚀失厚率数据进行增强，数据扩充对分析预测提供正向支持，生成数据不宜多于原始数据，防止扰乱训练样本分布，同时存在生成数据多样性受限的问题。

关键词：腐蚀失厚率；小样本；生成对抗网络；数据增强；降维分析；样本分布

中图分类号：TP399　　**文献标识码：**A　　**文章编号：**1672-9242(2023)01-0142-09

DOI：10.7643/ issn.1672-9242.2023.01.020

Corrosion Thickness Loss Rate Data Enhancement Based on a Small Sample of GAN

ZHOU Jun-yan, *WANG Jing-cheng*, *YANG Xiao-kui*, *SHU Chang*, *WANG Jin-mei*, *ZHANG Chen*

(Southwest Institute of Technology and Engineering, Chongqing 400039, China)

ABSTRACT: The work aims to conduct data enhancement on the corrosion thickness loss rate of small samples to achieve data expansion, improve the prediction accuracy of the subsequent analysis model, reduce the degree of overfitting and improve the generalization ability of the model. The Generative Adversarial Network (GAN) was used to expand the corrosion thickness loss rate data and make the data distribution more comprehensive. Dimensionality reduction visual analysis on the generated data was conducted. The distribution of generated data and original data samples was explored. The rationality of data enhancement was analyzed. In addition, the analysis and prediction ability and generalization ability were evaluated from the perspectives of multiple algorithm models and multiple evaluation indicators. The generated data filled in the weak link of the original data in the sample space distribution. After adding the generated data, the average MSE obtained by each machine learning algorithm model was 61.72% to 91.74% of the result without the generated data, and the Pearson average was 99.01% to 113.64 %. The prediction accuracy was improved. The results were more relevant. And the model generalization ability was enhanced. GAN can effectively enhance the corrosion thickness loss rate data of small samples. Data expansion provides positive support for

收稿日期：2021–11–17；**修订日期：**2021–12–28

Received：2021-11-17；**Revised：**2021-12-28

作者简介：周俊炎（1995—），男，硕士，工程师，主要研究方向为环境试验与观测。

Biography：ZHOU Jun-yan (1995-), Male, Master, Engineer, Research focus: environment test and observation research.

引文格式：周俊炎，王竟成，杨小奎，等. 基于 GAN 的小样本腐蚀失厚率数据增强方法[J]. 装备环境工程, 2023, 20(1): 142-150.

ZHOU Jun-yan , WANG Jing-cheng, YANG Xiao-kui, et al. Corrosion Thickness Loss Rate Data Enhancement Based on a Small Sample of GAN[J]. Equipment Environmental Engineering, 2023, 20(1): 142-150.

analysis and prediction. The generated data should not be more than the original data to prevent disturbing the distribution of training samples. At the same time, there are problems with limited diversity of generated data.

KEY WORDS: corrosion thickness loss rate; small sample; generative adversarial networks; data enhancement; dimensionality reduction analysis; sample distribution

腐蚀失厚率作为重要的环境效应数据，是金属板材最基本的腐蚀评价指标。如 El-Mahdy[1]以锌合金为出发点，研究腐蚀行为反映的环境污染程度，这些腐蚀数据具有重要的评估、经济价值。但由于金属服役环境复杂，导致环境影响因子不同；试验过程具有随机性、多变性、突变性及非线性等特点，导致存在各种不确定性因素；数据采集易受仪器、人员操作等噪声干扰，导致数据易出现奇异值；跨度时间漫长，导致数据易丢失、属性值不完整等问题。因此，最终可用于分析的完整可靠的腐蚀失厚率数据为小样本数据，迫切需要攫取数据潜在价值，实现腐蚀失厚率小样本数据集扩充，一定程度上代替传统的长时间跨度下环境试验收集数据方法，为后续腐蚀失厚率分析预测夯实数据基础。

目前，腐蚀失厚率分析预测主要的手段是根据样本分布与统计规律，直接套用各种数学函数定义腐蚀失厚率机理模型，实现小样本腐蚀失厚率数据分析预测，虽能达到一定精度，但泛化能力较弱，推广价值较低。如 Feliu 等[2]研究了大气中氯化物及二氧化硫对锌腐蚀的影响规律。王振尧等[3]及王光雍等[4]研究得出锌的大气腐蚀与试验时间为近似线性规律。从数据层面分析，其机理为假定了样本分布，但实际样本分布规律随机多变，尤其在小样本数据上，样本分布更加不稳定。

本文提出一种生成对抗网络（Generative Adversarial Networks，GAN）的模型用于小样本腐蚀失厚率数据扩充，提升数据价值密度[5]，以辅助后续分析预测。GAN 模型不要求样本分布，通过无监督学习的方式，使生成数据逼近真实样本分布[6]。本文通过调试 GAN 网络模型参数得到性能较好的腐蚀失厚率数据结果，通过 PCA 降维探究数据分布，并在多种机器学习算法上验证生成数据的可靠性及其对预测效果的提升。

1　腐蚀失厚率预测与 GAN 模型

1.1　腐蚀失厚率原始数据及预测分析指标

腐蚀失厚率原始数据囊括不同材料牌号、不同环境试验场景因素数据，其中不同环境试验场景通常以平均温度、相对湿度、降水量、日照时数等环境因素数据体现。本文腐蚀失厚率预测主要针对同材料牌号纯锌，在不同环境因素条件下腐蚀失厚率的数据预测。

锌在电位序中处于相对活性的位置，其腐蚀电位低于钢铁[7-8]，锌的大气腐蚀本质是薄液膜下锌金属的电化学腐蚀，其腐蚀行为具有较高的研究价值[9]。锌材料的腐蚀破坏产生严重的经济损失[10]，本文以腐蚀失厚率为依据之一。纯锌腐蚀失厚率数据示例见表 1，总数据量为 20 条，包含 12 个试验地点，记录其典型环境因素条目，并在大气环境暴露环境下测量 1 a 的腐蚀失厚率数据。

表 1　纯锌腐蚀失厚率数据示例

Tab.1 Example of pure zinc corrosion loss rate data

地点	腐蚀失厚率/($\mu m \cdot a^{-1}$)	平均温度/℃	年最高温度/℃	年最低温度/℃	平均相对湿度/%	最小相对湿度/%	日照时数/h	降水量/mm
万宁	6.49	24.41	31.19	17.65	85.99	32.55	179.41	162.69
武汉	1.96	16.82	27.29	8.47	71.58	24.42	211.65	99.7
库尔勒	0.59	13.03	19.77	5.73	50.38	26.4	55.08	0.57
拉萨	0.26	8.9	21.62	0.18	40.39	7.56	286.09	45.83

通过纯锌腐蚀失厚率数据训练得到预测模型，预测纯锌在其他各个试验地点的腐蚀失厚率，可通过全国乃至全球的环境因素数据直接预测得到纯锌的腐蚀失厚率，因此需要有效、合理利用这 20 条原始数据。由于某些站点进行了多样本测量，需要将数据顺序进行打乱，以 8∶2 的比例建立训练集与测试集。通常预测分析时采用交叉验证的方法，即多次重新取训练集与预测集进行结果分析，这样得出的结果更加可靠稳定，能减轻过拟合程度。但此方法中需要先训练 GAN 模型，再以 GAN 模型去生成数据，混合生成的增强数据与原始数据进行模型训练。考虑到交叉验证需要多个 GAN 模型，引入额外随机量，进行验证时变量过多，不具备理论说服力。因此，将训练集和测试集的数据固定，再采用降维可视化直接分析数据分布规律。

在后续腐蚀失厚率数据分析预测中，由于其主要数据表现形式为数值型，采用回归拟合的方式，拟采

用典型回归评价指标[11]。如赛晓勇等[12]用平均绝对误差、平均相对误差、误差平方和作为评价方法好坏的指标，以分析阳性检查效率。王雪等[13]用决定系数 R^2、均方根误差、验证误差、预测均方误差等作为红外光谱预测数据验证。参考董永权等[14]对相关系数的解释，误差绝对值也体现了统计特征[15]。本文使用均方差（MSE）、绝对值误差（MAE）、皮尔逊相关系数、相关指数 R^2 共 4 个指标作为回归结果评价、显著性检验手段。

1.2 生成对抗网络基本原理

生成对抗网络源于博弈论，以纳什均衡作为数学基础，是一种新的无监督学习算法框架[16]。二者博弈，使得二者不断以螺旋式动态提升，最终二者能力都得到显著增强。这种对抗的博弈方式，能够训练得出高质量的假数据，相比传统机器学习算法，泛化性能更好，生成模型能够学习到隐性的特征，具有更好的特征表达能力[17]。GAN 得到越来越多学者青睐，广泛运用在各个领域中，在计算机视觉领域最广泛、最成熟。一是由于图片由多个像素点组成，蕴含了大量隐藏特征属性；二是生成的图片具有直观属性，能够进行人工有效判断[18]。陈星宇等[19]将生成模型运用到图像视频显著性检测上。贝悦[20]等结合条件生成了对抗网络重构 HDR 图像，在计算机视觉方向成果显著。李凯伟等[21]利用 GAN 生成了情感对话内容。曹爽[22]基于生成对抗网络合成了表格数据。生成对抗网络逐渐普及到各个领域。

GAN 结构如图 1 所示，核心为 2 个网络结构，分别为生成器（Generator）和判别器（Discriminator），生成器的决策方向是尽可能生成逼近真实分布的假数据，让判别器无法识别生成的增强数据，而判别器决策方向是鉴别数据是否为真，尽可能区分真假数据。生成器和判别器形成对抗，从而不断迭代提升自身的判别或生成能力。当最终生成网络和判别网络的损失函数收敛时，一般情况下，此时判别网络能够一定程度上鉴真，但某些生成数据也会被判别为真数据，这就说明生成模型已经学习到真实样本的分布，并能够生成合理的假数据。

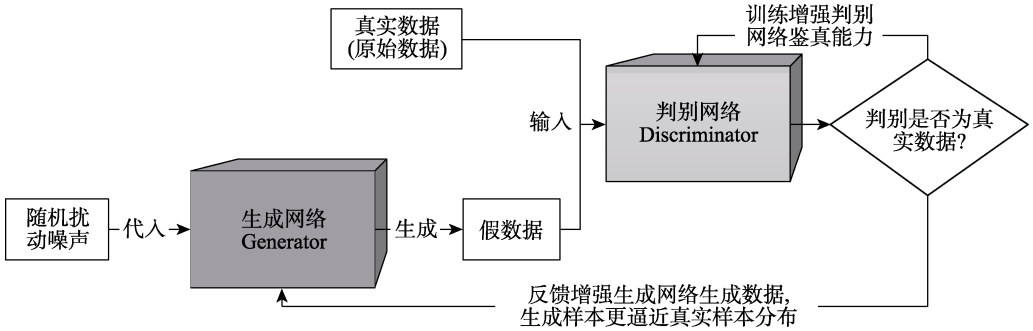


图 1 GAN 结构

Fig.1 Structure of generative adversarial networks (GAN)

在固定生成器 G 的情况下，需要最优化判别器 D，判别器迭代过程就是最小化交叉熵的过程，损失函数为：

$$Obj^{D}(\theta_D,\theta_G)=-\frac{1}{2}E_{x\sim p_{\text{real}}(x)}\lg D(x)-\frac{1}{2}E_{f\sim p_{\text{fake}}(f)}\lg\{1-D[G(f)]\} \tag{1}$$

式（1）中，真实数据 x 满足真实采样分布 $p_{\text{real}}(x)$，生成数据 f 满足先验分布 $p_{\text{fake}}(f)$，E 代表数据期望值。为达到需要，最小化式（1），在连续空间上有：

$$Obj^{D}(\theta_D,\theta_G)=-\frac{1}{2}\int_x p_{\text{real}}(x)\lg D(x)\mathrm{d}x-\frac{1}{2}\int_f p_{\text{fake}}(f)\lg\{1-D[G(f)]\}\mathrm{d}f=-\frac{1}{2}\int_x\left\{p_{\text{real}}(x)\lg D(x)+p_{\text{g}}(x)\lg\left[1-D(x)\right]\right\}\mathrm{d}x \tag{2}$$

式中：$p_{\text{g}}(x)$代表生成器的数据分布。对任意非零实数 m 和 n，且实数值 $y\in[0,1]$，表达式 $-m\lg y-n\lg(1-y)$ 在 $\frac{m}{m+n}$ 处取得最小值。由此可知，固定生成器 G 的情况下，目标函数（2）在式（3）处取得最小值。

$$D_G^{*}(x)=\frac{p_{\text{real}}(x)}{p_{\text{real}}(x)+p_{\text{g}}(x)} \tag{3}$$

由式（3）可知，目标函数取最小值即判别器的最优解，即 GAN 评估的对象为 2 个概率分布密度函数的比值。同时，判别器 D 的目的是当输入数据为真实数据 x 时，输出概率值 $D(x)$趋近于 1；当输入数据为生成数据 $G(f)$时，输出概率值 $D[G(f)]$趋近于 0。生成器 G 的目的是使 $D[G(f)]$趋近于 1，形成零和博弈，得到生成器 G 损失函数为 $Obj^{G}(\theta_G)=-Obj^{D}(\theta_D,\theta_G)$，最终变成一个极小–极大问题，GAN 目标函数等价于：

$$\min_G\max_D\{f(D,G)=E_{x\sim p_{\text{real}}(x)}\lg D(x)+E_{f\sim p_{\text{fake}}(f)}\lg\{1-D\left[G(f)\right]\}\} \tag{4}$$

GAN 训练过程即训练判别器不断最大化判别能力，同时不断训练生成器最小化判别能力。一般而言，

采用交替训练的方式，固定生成器 G，迭代优化判别器 D，然后固定判别器 D，迭代优化生成器 G，当生成器生成数据样本分布与原始数据样本分布对抗平衡时，达到全局最优解。

2　基于 GAN 的腐蚀失厚率生成式模型

GAN 普遍存在样本多样性较弱的问题[23]，但对于小样本腐蚀失厚率数据而言，若生成数据过于广泛，将导致实际物理意义不存在的问题。比如原始数据中有万宁、北京、武汉等地环境因素数据，如果生成差异性过大的数据，实际上不存在对应的真实地点，这样从机理层面无法解释，所以 GAN 生成偏向于真实数据的增强数据，使生成数据具有物理意义。因此，GAN 模型适用于腐蚀失厚率数据增强。

提出基于 GAN 的腐蚀失厚率生成式模型，并进行校验验证，流程如图 2 所示。首先是生成对抗网络主要组成部分的判别模型和生成模型，经参数调试，最适用于纯锌腐蚀失厚率数据分析的 GAN 模型网络结构为 4 层分类神经网络，判别器输入 8 个特征，中间层分别构建 16、256、64 个节点。二分类输出判别真假，使用 ReLU 作为激活函数，而生成器与判别器镜像对抗，输入为 1 个特征，中间层构建 64、256、16 个网络节点，最后输出 8 个特征，即假数据。利用小样本的腐蚀失厚率数据进行 GAN 训练，并保存生成网络模型。生成网络模型生成假数据，与真数据进行 PCA 降维可视化分析，校验数据是否分布合理。同时对比真数据、混合真数据和假数据在不同算法的预测结果，输出得到各类评价指标，以此验证基于 GAN 进行数据增强后是否对预测精度提升提供支持。

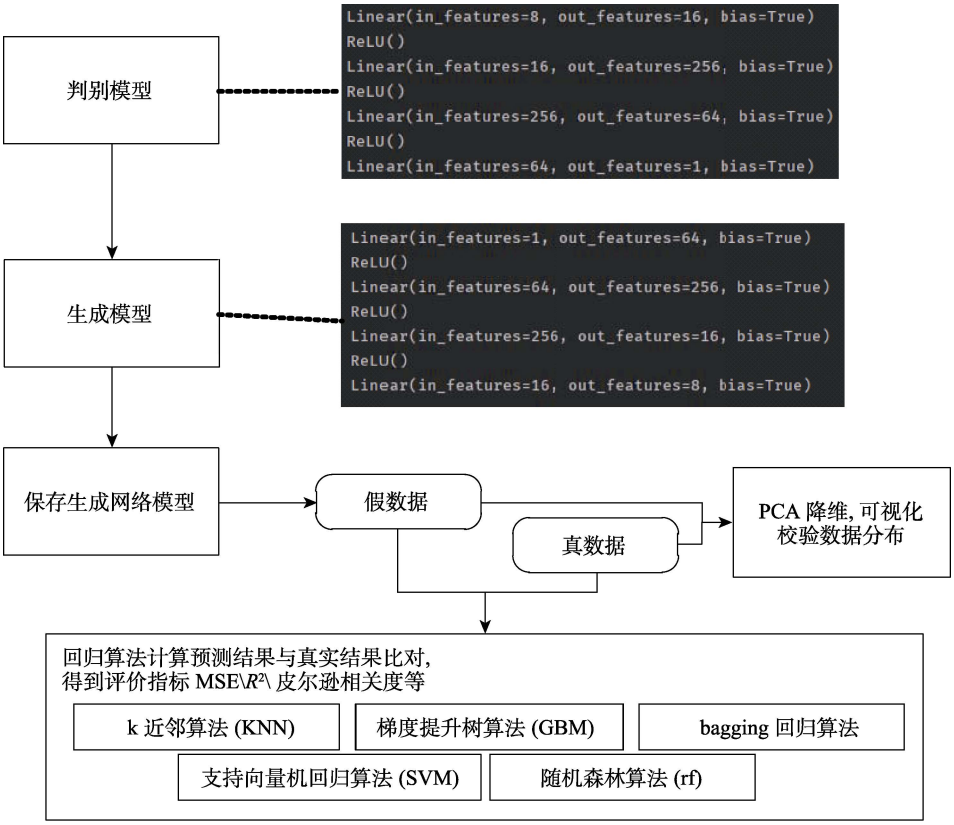


图 2　腐蚀失厚率生成式模型流程

Fig.2 Generative model process of corrosion loss rate

保存收敛的生成网络模型，要求腐蚀失厚率 GAN 模型达到收敛。输出判别器与生成器的损失函数值，得到如图 3 所示损失函数曲线。图 3 中，迭代次数指损失输出次数，采取措施是前 200 代每 10 次输出 1 次损失函数，之后每 50 代输出 1 次损失函数，所以 600 多次迭代次数对应实际 30 000 次循环。判别器与生成器损失在初期 100 次（即实际 4 200 代）以内波动非常大，150 次迭代次数（即实际 6 700 代）后缓慢收敛，最终取 30 000 次循环结果作为收敛结果。最终判别器损失收敛于 0.7 左右，生成器损失收敛于 2.3 左右，所以判别器基本稳定收敛，生成器由于数据量较少，只能达到基本收敛。另外，GAN 的收敛不能只取决于判别器与生成器的收敛，同时输出真假数据在判别器的概率，当真假数据概率基本持平时，模型达到收敛。当 GAN 模型达到收敛时，将生成模型参数、网络结构保存，以备后续生成假数据、分析使用。

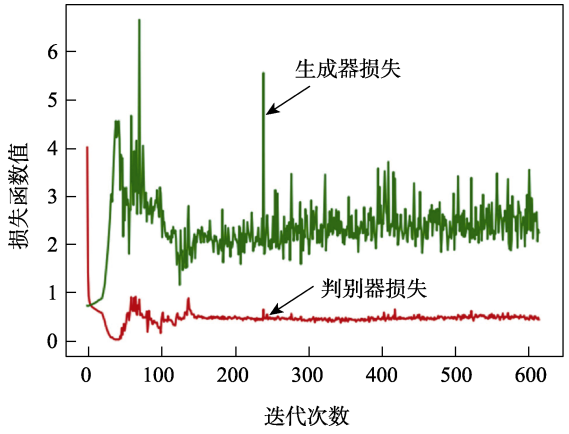


图 3　腐蚀失厚率 GAN 模型损失函数

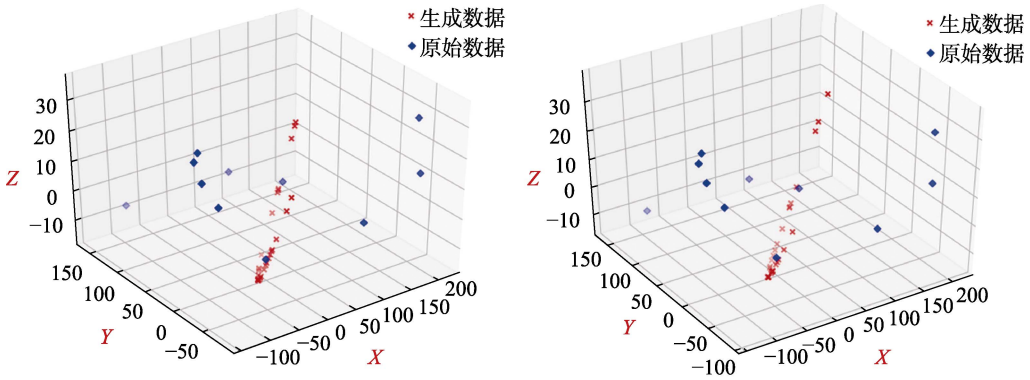
Fig.3 GAN model loss function of corrosion loss rate

3　结果及分析

3.1　PCA 校验数据分布

纯锌腐蚀失厚率数据为 7 个特征列与 1 个标签列，共 8 维数据，无法直观分析其分布规律，需要将其降维到三维及其以下可视化数据特征。这里采用主成分分析技术(Principal Components Analysis，PCA)，利用方差信息进行线性变换投影降维，压缩数据空间，将高维度的纯锌腐蚀失厚率数据在三维空间中直观展示出来[24]。

如图 4a 所示的不带 label 列的 PCA 三维降维结果，指只使用生成数据的特征列进行降维。从三维降维可视化结果可得，原始的 20 条小样本数据散乱地分布在样本空间中，并且在 X、Y、Z 值都较大或都较小的情况时，数据不存在，生成模型主要填补了此种情况下的样本分布，使样本分布更加完整。带 label 列的 PCA 三维降维结果与不带 label 列的降维结果基本一致，主要补齐样本在某些情况下的分布，如图 4b 所示。GAN 生成的数据特征主要集中在非 label 列，即特征列上。因此，GAN 生成的腐蚀失厚率假数据的样本分布合理，可支撑后续分析研究，但生成数据多样性不够丰富。由可视化结果可知，主要补齐的数据沿 2 条直线分布（低维线性分布，在原始高维分布一般不为线性分布），存在 GAN 典型的模式坍缩问题。



a 不带 label 列　　b 带 label 列

图 4　PCA 三维降维结果

Fig.4 PCA 3D dimensionality reduction result: a) without label column; b) with label column

3.2　预测结果分析

为了验证生成数据是否可以提升预测的精度，增加模型的泛化能力，采用极端随机树回归算法（ETR）、梯度提升树算法（GBM）、bagging 回归算法、支持向量机回归算法（SVM）、随机森林回归算法（RF）共 5 种机器学习算法，分别使用原始数据、混合原始数据加部分生成数据 2 种方式进行模型训练。原始数据通过 8∶2 的比例划分训练集与测试集，即使用 16 条数据进行训练，而加入生成数据的策略为使用 16 条原始数据，并加上 4 条生成数据，保证原始数据权重，防止训练数据样本分布过于偏向生成数据样本分布。最后以 4 条测试集数据实际值与预测值进行均方差（MSE）、绝对值误差（MAE）、相关指数 R^2、皮尔逊相关系数 4 个评价指标来验证结果。

由于多数回归模型都具有一定随机性，比如梯度提升树，该模型不断随机抓取数据进行梯度提升以达最优解，因此每次训练模型得到的结果存在一定波动性。鉴于此，采用多次训练模型，并取统计直方图的方法进行效果校验，取 10 000 次循环输出统计直方图。以随机森林为例，各评价指标的统计直方图见图 5。对于均方差和绝对值误差而言，加入生成数据训练的模型结果均值更小，模型准确度更高，同时方差更大，说明通过加入样本扩充分布后随机性得到一定提升，过拟合程度相对降低。对于皮尔逊相关系数和相关指数 R^2 而言，加入生成数据训练的模型结果均值更大，说明预测值与实际值关联性更强，方差更大，增添随机性，过拟合程度相对降低。

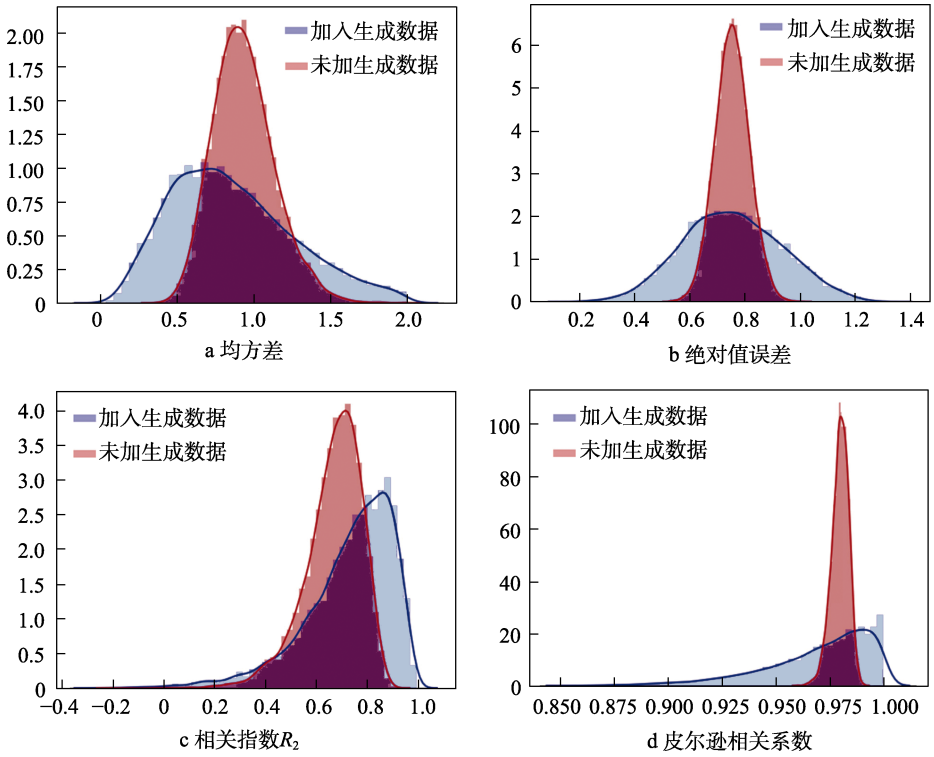


图 5　数据增强前后随机森林评价指标

Fig.5 Ramd evaluation indicators before and after data enhancement: a) mean square deviation; b) absolute value error; c) correlation index R_2; d) Pearson correlation coefficient

参考袁培等[25]从多维、多源数据进行多尺度分析，本文通过 5 种回归算法分别对使用原始数据、原始数据加部分生成数据 2 种方式的训练结果进行评价指标分析。由于每种算法多次测算，以多次测算的均值方差来代替直方图表达，最终统计得到数据增强前后各模型评价指标数值，见表 2。特别地，对于 SVM 回归而言，由于 SVM 是一种数值解法，若输入一样，每次结果一致，不存在随机过程，所以未加入随机生成数据时，使用同样的 16 个训练数据得到的结果完全相同，体现为方差等于 0。

表 2　数据增强前后各模型评价指标

Tab.2 Evaluation indicators of each model before and after data enhancement

算法模型	是否加入生成数据	MSE		MAE		R^2		皮尔逊	
		均值	方差	均值	方差	均值	方差	均值	方差
ETR 回归	是	0.257 0	0.034 346 64	0.275 3	0.012 209 48	0.960 0	0.000 833 35	0.987 0	0.000 092 77
	否	0.310 8	0.664 239 27	0.280 5	0.023 294 19	0.955 0	0.007 697 89	0.984 7	0.001 346 68
梯度提升树	是	0.138 1	0.013 055 69	0.209 5	0.007 784 95	0.978 5	0.000 327 73	0.993 3	0.000 035 15
	否	0.223 1	0.000 695 62	0.283 0	0.000 197 87	0.965 1	0.000 035 22	0.989 0	0.000 002 00
bagging 回归	是	0.840 9	0.232 269 85	0.720 5	0.051 070 02	0.733 3	0.053 180 27	0.970 0	0.000 756 37
	否	1.073 5	0.445 743 49	0.756 0	0.037 014 58	0.477 3	0.443 642 14	0.975 2	0.000 310 59
SVM 回归	是	4.398 1	0.622 668 95	1.264 6	0.007 770 33	–7.592 8	74.447 221 59	0.846 2	0.008 184 08
	否	5.251 8	0	1.383 5	0	–10.191 7	0	0.740 6	0
随机森林回归	是	0.863 6	0.160 160 96	0.756 0	0.032 250 92	0.728 9	0.031 138 67	0.970 1	0.000 617 81
	否	0.939 6	0.039 875 37	0.755 2	0.003 848 75	0.666 6	0.012 864 49	0.979 9	0.000 015 94

通过统计计算，加入生成数据后的 MSE 均值是未加入生成数据 MSE 均值的 61.72%~91.74%，皮尔逊均值为 99.01%~113.64%。MSE 综合衡量偏差与方差，模型精确度提升，皮尔逊均值衡量预测值和实际值之间相关性，结果关联度更高。

根据表 2 得到综合图示，如图 6 所示，直观展示各算法模型对各评价指标的数值结果，ETR、gbm、bagging、SVM、RF 分别指极端回归树、梯度提升树、

bagging 回归、支持向量机回归、随机森林回归算法模型，“是”与“否”代表是否加入生成数据，如“ETR-是”指加入生成数据的极端回归树算法结果。图 6 中负指标缩小为 10%处理，且未展示方差。首先因为方差数值量级差异较大，难以直观展示，其次方差体现随机性，而随机性可通过 PCA 降维分析推导或者方差计算得出。由图 6 可直观得出，加入生成数据后，MSE、MAE 均值降低，R^2、皮尔逊均值增大。

为了探究生成数据数量对分析预测结果的影响，这里使用梯度提升树算法（GBM）为基础，测试取不同量生成数据 gbm 算法结果，见表 3。可见，随着生成数据取用数量的增多，MSE 逐渐趋小，但同时存在随机性减小、过拟合程度增大的问题。从 MSE 方差角度看，加入生成数据后，方差增加几十倍，生成数据取用数为 4 左右时达到峰值，说明此时随机性更好，样本分布更加全面。因此，对于腐蚀失厚率数据而言，纳入训练的原始数据 16 条，再加入 4 条生成数据时，分析预测精度高，随机性更好，样本分布更全面。将实验结果绘制成折线图，如图 7 所示。生成数据只要不超过原始数据数量的 100%，分析预测结果较为可观。

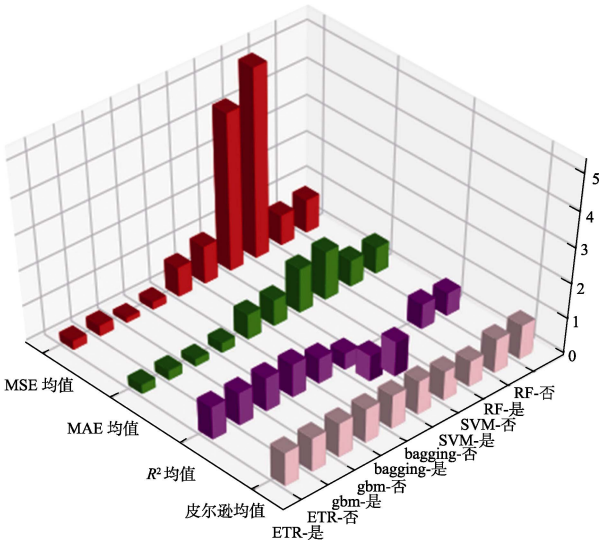


图 6　数据增强前后各算法评价指标

Fig.6 Diagrams of evaluation indicators for each algorithm before and after data enhancement

表 3　取不同量生成数据 gbm 算法结果

Tab.3 Results of the gbm algorithm for different amounts of generated data

取生成数据数量	MSE		MAE		R^2		皮尔逊	
	均值	方差	均值	方差	均值	方差	均值	方差
0	0.223 1	0.000 695 62	0.283 0	0.000 197 87	0.965 1	0.000 035 22	0.989 0	0.000 002 00
4	0.138 1	0.013 055 69	0.209 5	0.007 784 95	0.978 5	0.000 327 73	0.993 3	0.000 035 15
8	0.148 4	0.012 709 41	0.221 2	0.007 098 69	0.976 5	0.000 319 47	0.992 6	0.000 034 57
12	0.144 9	0.011773 17	0.217 8	0.006 794 65	0.977 4	0.000 295 40	0.993 0	0.000 032 00
16	0.145 3	0.011 856 38	0.218 0	0.006 832 39	0.977 4	0.000 297 42	0.992 9	0.000 032 23
32	0.187 0	0.012 031 07	0.249 3	0.005 755 31	0.970 7	0.000 303 60	0.990 8	0.000 033 32
64	0.186 8	0.012 018 90	0.249 3	0.005 688 17	0.970 8	0.000 303 36	0.990 8	0.000 033 31

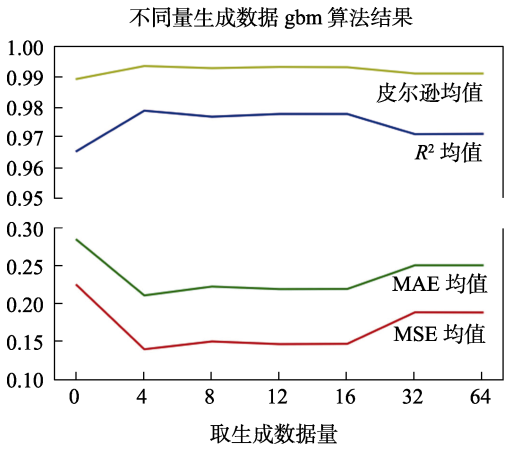


图 7　不同量生成数据 gbm 算法结果展示

Fig.7 Display of results of the gbm algorithm for different amounts of generated data

4　结论

1）综合考虑腐蚀失厚率小样本数据特点，提出适用的 GAN 算法模型，调整 4 层网络结构及参数。

2）使用多算法模型、多个评价指标对 GAN 数据增强结果进行验证，结果表明，GAN 数据扩充能有效对小样本腐蚀失厚率分析预测提供可靠支持。

3）对比取不同数量的增强数据分析预测结果，采用适中数量的生成数据才能得到最佳的分析预测结果。试验证明，生成数据小于原始数据数量时，腐蚀失厚率分析效果较好，不会扰乱样本分布。

4）生成数据存在多样性不够充分问题，GAN 存在模式坍缩问题，后续可探究 WGAN 等更复杂的网络结构，或者通过数据清洗等方法优化样本分布，以解决存在的问题。

参考文献：

[1]　EL-MAHDY G A. Advanced Laboratory Study on the Atmospheric Corrosion of Zinc under Thin Electrolyte Layers[J]. Corrosion, 2003, 59(6): 505-510.

[2] FELIU S, MORCILLO M, FELIU S Jr. The Prediction of Atmospheric Corrosion from Meteorological and Pollution Parameters—I. Annual Corrosion[J]. Corrosion Science, 1993, 34(3): 403-414.

[3] 王振尧, 于国才, 韩薇. 我国若干典型大气环境中的锌腐蚀[J]. 腐蚀科学与防护技术, 2003, 15(4): 191-195.
WANG Zhen-yao, YU Guo-cai, HAN Wei. Atmospheric Corrosion Performance of Zinc at Several Selected Test Sits in China[J]. Corrosion Science and Technology Protection, 2003, 15(4): 191-195.

[4] 王光雍, 王海江, 李兴濂, 等. 自然环境的腐蚀与防护[M]. 北京: 化学工业出版社, 1996.
WANG Guang-yong, ANG hai-jiang, LI Xing-lian, et al. Corrosion and Protection of the Natural Environment[M]. Beijing: Chemical Industry Publishing House, 1996.

[5] 尚宇炜, 马钊, 彭晨阳, 等. 内嵌专业知识和经验的机器学习方法探索(一): 引导学习的提出与理论基础[J]. 中国电机工程学报, 2017, 37(19): 5560-5571.
SHANG Yu-wei, MA Zhao, PENG Chen-yang, et al. Study of a Novel Machine Learning Method Embedding Expertise Part Ⅰ: Proposals and Fundamentals of Guiding Learning[J]. Proceedings of the CSEE, 2017, 37(19): 5560-5571.

[6] 杨懿男, 齐林海, 王红, 等. 基于生成对抗网络的小样本数据生成技术研究[J]. 电力建设, 2019, 40(5): 71-77.
YANG Yi-nan, QI Lin-hai, WANG Hong, et al. Research on Generation Technology of Small Sample Data Based on Generative Adversarial Network[J]. Electric Power Construction, 2019, 40(5): 71-77.

[7] 郝显赫, 王振尧, 汪川. 锌在辽宁红沿河核电站的大气腐蚀研究[J]. 装备环境工程, 2012, 9(3): 108-110.
HAO Xian-he, WANG Zhen-yao, WANG Chuan. Atmospheric Corrosion of Zinc at Hongyanhe Nuclear Power Station[J]. Equipment Environmental Engineering, 2012, 9(3): 108-110.

[8] 周学杰, 张三平, 郑鹏华, 等. 纯锌在水环境中腐蚀行为[J]. 装备环境工程, 2008, 5(5): 9-12.
ZHOU Xue-jie, ZHANG San-ping, ZHENG Peng-hua, et al. Corrosion Behavior of Pure Zn in Water Environment[J]. Equipment Environmental Engineering, 2008, 5(5): 9-12.

[9] 叶堤. 重庆市大气污染对锌材料腐蚀的经济损失分析[J]. 装备环境工程, 2007, 4(1): 21-24.
YE Di. Economic Loss Estimates of Zinc Corrosion by Acid Deposition in Chongqing[J]. Equipment Environmental Engineering, 2007, 4(1): 21-24.

[10] 朱志平, 左羡第, 银朝晖. 锌在模拟工业大气环境下的腐蚀行为研究[J]. 装备环境工程, 2015, 12(4): 1-5.
ZHU Zhi-ping, ZUO Xian-di, YIN Zhao-hui. Zinc Corrosion Behavior in Simulated Industrial Atmospheric Environment[J]. Equipment Environmental Engineering, 2015, 12(4): 1-5.

[11] 王振杰, 姚吉利. 广义测量平差分类[J]. 淄博学院学报(自然科学与工程版), 2001(1): 62-64.
WANG Zhen-jie, YAO Ji-li. The Classification of General Surveying Adjustment[J]. Journal of Zibo University, 2001(1): 62-64.

[12] 赛晓勇, 邢秦菊, 孟定茹, 等. 五种预测方法在退田还湖区血吸虫病发病的拟合效果评价[J]. 第四军医大学学报, 2006(17): 1603-1605.
SAI Xiao-yong, XING Qin-ju, MENG Ding-ru, et al. Comparison of Predicting Effect of Schistosomiasis Prevalence by 5 Statistical Models in the Areas of “Breaking Dikes or Opening Sluice for Water Store” in Dongting Lake[J]. Journal of the Fourth Military Medical University, 2006(17): 1603-1605.

[13] 王雪, 马铁民, 杨涛, 等. 基于近红外光谱的灌浆期玉米籽粒水分小样本定量分析[J]. 农业工程学报, 2018, 34(13): 203-210.
WANG Xue, MA Tie-min, YANG Tao, et al. Moisture Quantitative Analysis with Small Sample Set of Maize Grain in Filling Stage Based on near Infrared Spectroscopy[J]. Transactions of the Chinese Society of Agricultural Engineering, 2018, 34(13): 203-210.

[14] 董永权, 王占民. 关于相关系数 ρ 的几点注释[J]. 大学数学, 2008, 24(2): 182-186.
DONG Yong-quan, WANG Zhan-min. Some Notes about Correiation Coefficient ρ[J]. College Mathematics, 2008, 24(2): 182-186.

[15] 丁勇. 误差绝对值的统计特征和应用[J]. 数理统计与管理, 2016, 35(1): 39-46.
DING Yong. The Statistical Characteristic and Application of the Error Absolute Value[J]. Journal of Applied Statistics and Management, 2016, 35(1): 39-46.

[16] 王坤峰, 苟超, 段艳杰, 等. 生成式对抗网络 GAN 的研究进展与展望[J]. 自动化学报, 2017, 43(3): 321-332.
WANG Kun-feng, GOU Chao, DUAN Yan-jie, et al. Generative Adversarial Networks: The State of the Art and beyond[J]. Acta Automatica Sinica, 2017, 43(3): 321-332.

[17] 梁俊杰, 韦舰晶, 蒋正锋. 生成对抗网络 GAN 综述[J]. 计算机科学与探索, 2020, 14(1): 1-17.
LIANG Jun-jie, WEI Jian-jing, JIANG Zheng-feng. Generative Adversarial Networks GAN Overview[J]. Journal of Frontiers of Computer Science and Technology, 2020, 14(1): 1-17.

[18] 陈亮, 吴攀, 刘韵婷, 等. 生成对抗网络 GAN 的发展与最新应用[J]. 电子测量与仪器学报, 2020, 34(6): 70-78.
CHEN Liang, WU Pan, LIU Yun-ting, et al. Development and Application of the Latest Generation Against the Network of GAN[J]. Journal of Electronic Measurement and Instrumentation, 2020, 34(6): 70-78.

[19] 陈星宇, 叶锋, 黄添强, 等. 融合小型深度生成模型的显著性检测[J]. 电子学报, 2021, 49(4): 768-774.
CHEN Xing-yu, YE Feng, HUANG Tian-qiang, et al. Saliency Detection Combined with Small-Scale Deep Generation Model[J]. Acta Electronica Sinica, 2021, 49(4):

768-774.

[20] 贝悦, 王琦, 程志鹏, 等. 基于条件生成对抗网络的 HDR 图像生成方法[J]. 北京航空航天大学学报, 2022, 48(1): 45-52.
BEI Yue, WANG Qi, CHENG Zhi-peng, et al. HDR Image Generation Method Based on Conditional Generative Adversarial Network[J]. Journal of Beijing University of Aeronautics and Astronautics, 2022, 48(1): 45-52.

[21] 李凯伟, 马力. 基于生成对抗网络的情感对话回复生成[J/OL]. 计算机工程与应用, 2021: 1-8. (2021-04-22). https://kns.cnki.net/kcms/detail/11.2127.TP.20210422.1328.011.html.
LI Kai-wei, MA Li. Emotional Dialogue Response Generation Based on Generative Adversarial Network[J/OL]. Computer Engineering and Applications, 2021: 1-8. (2021-04-22). https://kns.cnki.net/kcms/detail/11.2127.TP.20210422.1328.011.html.

[22] 曹爽. SCGAN: 合成单类别表格数据的生成对抗网络[J]. 计算机时代, 2021(4): 25-27.
CAO Shuang. SCGAN: A Generative Adversarial Network for Single Category Tabular Data Synthesis[J]. Computer Era, 2021(4): 25-27.

[23] RADFORD A, METZ L, CHINTALA S. Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks[J]. Computer Science, 2015, 1511: 06434.

[24] 陈佩. 主成分分析法研究及其在特征提取中的应用[D]. 西安: 陕西师范大学, 2014: 8-15.
CHEN Pei. Research on Principal Component Analysis and Its Application in Feature Extraction[D]. Xi'an: Shaanxi Normal University, 2014: 8-15.

[25] 袁培, 王舶仲, 毛文奇, 等. 基于多重生成对抗网络的智能开关设备状态感知与诊断研究[J]. 电力系统保护与控制, 2021, 49(6): 67-75.
YUAN Pei, WANG Bo-zhong, MAO Wen-qi, et al. Research on State Perception and Diagnosis of Intelligent Switches Based on Triple Generative Adversarial Networks[J]. Power System Protection and Control, 2021, 49(6): 67-75.

责任编辑：刘世忠