

基于半监督生成对抗网络的水泥熟料游离氧化钙数据增强方法及其有效性评价

姚歆玥, 任雪红, 叶家元, 翟牧楠, 曹傅荔, 张文生, 张洪滔

(中国建筑材料科学研究总院有限公司, 北京 100024)

摘 要: 高效、准确地实时预测游离氧化钙(f-CaO)含量对于实现水泥质量的稳定控制至关重要。针对生产过程中采样周期不一致引发的标签数据缺失、f-CaO 含量预测精度低的问题, 本研究提出并采用了半监督生成对抗网络(SSGAN)的方法, 对 f-CaO 数据集进行数据增强, 并借助门控递归单元模型, 对输入方式及编码器等关键模块进行了消融实验, 评估了其有效性。结果表明, SSGAN 方法可以生成与真实数据分布高度一致的虚拟标签数据, 显著扩充了训练样本的多样性和代表性, 并使得下游预测任务的平均绝对误差、均方误差、均方根误差分别由 0.24、0.08、0.29 降至 0.19、0.06、0.24, 证实了 SSGAN 策略在降低预测误差和提升模型鲁棒性方面的显著作用, 为材料数据增强研究提供了可借鉴的方法。

关键词: 游离氧化钙; 半监督生成对抗网络; 时间序列; 数据增强; 回归预测; 消融实验

中图分类号: TQ172.11 **文献标志码:** A **文章编号:** 0454-5648(2025)11-3127-09

网络出版时间: 2025-09-29



水泥熟料中的游离氧化钙(f-CaO)含量是衡量其煅烧质量及工艺稳定性的核心指标, 直接影响水泥的安定性和力学性能^[1-2]。传统检测方法依赖人工进行离线化学分析或X射线衍射分析, 其检测周期长、滞后性高且操作复杂, 难以及时反馈生产参数调整需求^[3-4]。随着工业大数据与人工智能技术的迅速发展, 机器学习方法因其在解决高度非线性和大时滞问题上的潜力, 逐渐成为f-CaO含量预测研究领域的热点。

早期研究主要采用线性回归^[5]、支持向量机^[6]等浅层模型, 并结合优化算法或集成学习的方法提升预测性能^[7], 但大多依赖静态时间点匹配策略, 难以反映生产过程的时序动态特征。随着分布式控制系统(DCS)在水泥生产中的普及, 高频采样的工艺参数数据得以有效获取, 从而推动了深度学习技术的应用^[8-9]。Zhou等^[10]将工艺参数转换为二维过程数据块, 并引入模型级权重更新的域自适应方法, 对卷积神经网络进行优化, 实现了f-CaO含量的有效预测。Zhao等^[11]则利用人工神经网络(ANN)从多变

量时间序列中提取特征, 构建了基于多元时间序列分析和卷积神经网络的f-CaO含量预测模型, 具有较高的准确率和泛化能力。

尽管已有方法在时序性建模上取得进展, 但实际生产中工艺参数数据采样频率约为1 s, 而f-CaO含量的检测周期通常为1 h, 导致标签数据严重稀缺。这种异步采样与数据异构性限制了监督学习模型训练的数据规模和多样性, 从而降低了模型的预测精度和泛化能力^[12]。近年来, 生成对抗网络(GAN)的提出为数据增强和数据生成提供了一种全新的思路^[13-14], 通过生成器与判别器之间的对抗训练, 学习数据分布并生成高质量样本^[15-16]。GAN方法能够充分利用大量未标记数据进行补全, 提升数据多样性和模型泛化能力已逐渐成为学术界和工业场景下的研究热点^[17-19], 为解决水泥熟料生产过程中采样频率不一致导致的未标记数据问题提供了新方法, 但目前少见有应用GAN预测熟料f-CaO含量的研究。

基于此, 本研究提出了一种基于半监督对抗生

收稿日期: 2025-04-03。 修订日期: 2025-04-22。

基金项目: 国家自然科学基金双碳专项(52341202)。

第一作者: 姚歆玥(1999), 女, 硕士研究生。

通信作者: 任雪红(1984), 女, 博士, 教授。

Received date: 2025-04-03. Revised date: 2025-04-22.

First author: YAO Xinyue (1999-), female, Master candidate.

E-mail: yaoxinyuewhut@163.com

Correspondent author: REN Xuehong (1984-), female, Ph.D., Professor.

E-mail: renxue968@163.com

成网络(SSGAN)的方法, 扩充训练集的多样性及其代表性, 并通过其对下游f-CaO含量预测任务精度的影响和消融实验, 探究了GAN在处理工艺参数与离线检测数据之间异步采样问题上的应用可行性。

1 实验

1.1 数据集构建

从某水泥企业的硅酸盐水泥熟料产线连续收集了6个月共计116 d的DCS系统数据, 以及与其对应熟料的f-CaO含量结果。经过统计分析, 样本中f-CaO含量小于1.5%的合格样本占比为84.24%, 占据绝大多数; 同时, 数据中涵盖了少量f-CaO含量超过1.5%的样本离散点(约占15.76%), 主要是由于生料均化不充分或熟料煅烧温度波动等异常因素引起的^[20]。样本收集时间较长、数据分布范围较广, 有利于提升构建模型的泛化性能。

结合时间窗口划分策略, 在DCS系统数据的基础上构建了一个基于多元时间序列的原始数据集。考虑到熟料从进入生料提升机、经过预热器、回转窑煅烧、最终在篦冷机完成冷却的整个烧成流程大约需时60 min左右, 因此以60 min作为一个时间

窗口。每个样本由该时间段内连续采集的工艺参数序列构成输入特征, 反映了烧成过程的动态演化; 同时, 将窗口结束时刻所对应的f-CaO检测值作为标签, 实现了工艺参数序列与质量指标之间的时间对齐, 构成多元时间序列与对应标签值的原始数据集。

在实际生产过程中, DCS系统通过大量传感器来自动监测工艺参数, 因此系统中涉及大量过程参数。如果选择过多的特征参数作为模型的输入, 会大大增加模型训练和计算的时间, 不能满足实际应用的实时性要求。因此, 本研究针对影响熟料煅烧的物料量、煤量、气流等关键工艺参数, 并结合相关研究^[21–22], 从180个工艺参数中筛选出了生料提升机电流、头煤瞬时量、窑转速、分解炉出口温度、出篦冷机熟料温度等16个与熟料煅烧的热工状态密切相关、对f-CaO含量具有显著影响的主要工艺参数, 见表1所示, 以期降低构建模型的复杂度, 提高模型的泛化能力和预测精度。其中, 部分参数整体缺失率较高, 为了避免对预测模型的精度产生影响, 对数据集进行了数据清洗及均值滤波、标准化、时间窗口划分等预处理操作, 构建了一个含有2196组数据的工艺参数数据集。

表1 工艺参数原始数据集的变量统计结果

Table 1 Statistical results of the variables in the original process parameters dataset

Variable	Mean	Std	Min	Max	Category
Raw meal hoist current 1#	126.26	5.29	74.99	147.75	Input
Raw meal hoist current 2#	129.64	5.60	79.22	152.29	Input
High-temperature fan current	224.77	12.00	0.26	268.67	Input
Head coal	11.91	0.66	0.01	13.86	Input
Tailing coal	19.60	2.37	0.00	35.19	Input
Rotation speed of the kiln	4.12	0.18	0.00	4.39	Input
CO ₂ content of C1	1677.44	585.41	0.00	23702.98	Input
C5A temperature	886.81	13.91	0.00	1142.97	Input
C5B temperature	842.37	31.16	0.00	1057.03	Input
Tertiary air temperature	1013.78	79.38	0.00	1292.42	Input
Secondary air temperature	1085.99	76.61	0.00	1422.18	Input
Precalciner outlet temperature	888.68	22.36	663.31	1205.90	Input
Clinker temperature	117.28	47.46	0.26	531.53	Input
flue chamber temperature	1090.70	49.25	0.00	1341.38	Input
O ₂ content of C1	3.74	4.43	0.00	26.36	Input
NO ₂ content of C1	60.92	58.43	0.00	3070.31	Input
f-CaO content	1.27	0.77	0.05	13.00	Output

1.2 数据增强模型构建

本工作提出了一种基于SSGAN的数据增强模

型, 通过生成与实际数据分布相似的虚拟数据, 从而扩充训练集的样本量, 并有效提高模型在缺乏标

记数据情况下的学习能力。生成的虚拟数据不仅能够捕捉样本数据中的高维特征,还能够模拟出与实际生产过程高度一致的情况。同时,为了缓解传统 GAN 框架中易出现的模式崩溃问题,进一步采用了 Wasserstein GAN with Gradient Penalty(WGAN-GP)的训练策略,通过优化 Wasserstein 距离并对损失

函数的梯度施加惩罚,确保网络满足 Lipschitz 连续性,有效缓解模式崩溃,加速收敛过程,提供平滑稳定的对抗梯度,推动生成器生成更贴近真实数据分布的样本。图 1 展示了数据增强模型的总体架构,主要分为编码器、生成器网络和判别器网络 3 个部分。

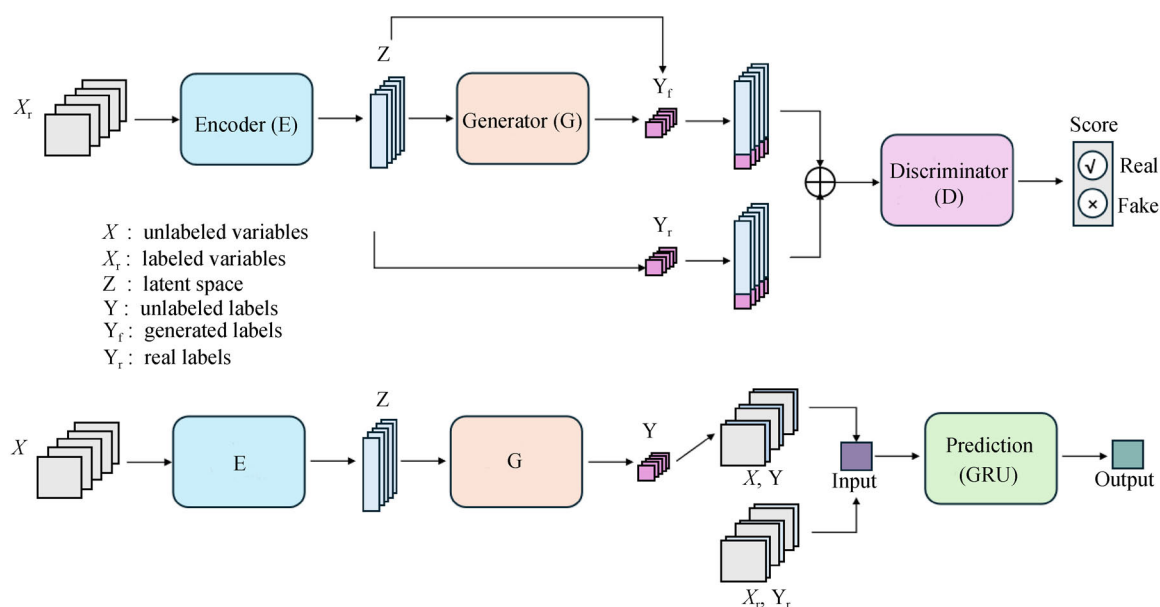


图 1 基于 SSGAN 的数据增强模型整体流程图

Fig. 1 Overall flowchart of data enhancement model based on SSGAN

与传统仅依赖随机噪声输入的方式不同,本研究首先通过编码器对少量带有真实标签的工艺参数数据矩阵进行特征提取,并以此作为条件输入。生成器则通过对抗训练来学习真实标签分布的潜在特征,并利用未标记的工艺参数数据生成对应的标签值。为了充分提取高维时序特征和时间尺度特征,编码器结构融合了多层卷积与门控递归单元(GRU),并在生成器网络中引入反卷积操作,从而更好地捕捉数据在时序维度上的复杂关联。判别器网络主要承担了“真伪判别”的任务:其通过对真实标签与生成标签进行区分,指导生成器不断优化其生成策略。随着对抗训练的迭代进行,生成器能够生成与实际生产过程高度相似的合成标签数据,判别器则在学习真实数据分布与标签特征方面持续提升。

在完成对抗训练后,生成器输出的虚拟样本将与真实样本相结合,构建出更具多样性与代表性的训练数据集,并将该数据集输入到GRU预测模型进行训练。通过上述编码器、生成器网络、判别器网

络的协同设计,整体框架不仅能缓解标记数据不足的问题,还能在工艺条件分析与熟料性能预测中取得更高的精度与可靠性。

1.3 参数设置

SSGAN数据增强模型的开发与实现是在Python3.9环境下进行的,依托于Pandas、Numpy、torch.nn以及其他基于PyTorch框架的相关包构建模型。编码器首先通过两层卷积网络(分别采用32和64个通道,卷积核大小均为3且采用填充1)提取局部特征,随后利用GRU将特征降维至32维;生成器则基于转置卷积结构对工艺特征进行映射,生成对应的单维标签值,而判别器通过先将标签值与编码器输出拼接,经过全连接层及两层卷积网络进行判别,从而对生成效果进行评估。在模型训练过程中,生成器与编码器共享相同优化器,均采用Adam优化器, Batch size设置为32、学习率0.000 015, β_1 设为0.5, β_2 为0.999。在判别器的训练过程中,采用了梯度惩罚机制, λ 设置为0.3,以满足Lipschitz连续性约束,并通过每5个批次更新一次生成器的策略,确保

训练过程的稳定性和生成样本的质量。

1.4 评价指标

所有训练在工艺参数数据集上进行, 并将数据集按8:2的比例划分为相同的训练集和测试集, 以确保模型评估的准确性和有效性。使用平均绝对误差(MAE)、均方误差(MSE)、均方根误差(RMSE)作为指标来评估回归模型的预测精度, 其公式定义如式(1)~式(3)所示。

$$MAE = \frac{1}{n} \sum_{i=1}^n (T - Y) \quad (1)$$

$$MSE = \frac{1}{n} \sum_{i=1}^n (T - Y)^2 \quad (2)$$

$$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{n} \sum_{i=1}^n (T - Y)^2} \quad (3)$$

式中: T 和 Y 分别为实际结果和预测结果。作为评价模型预测精度的指标, 通常MAE、MSE、RMSE越接近于0, 预测精度越高。

2 结果与讨论

2.1 模型收敛性分析

在SSGAN的训练过程中, 模型的收敛性是评估其性能和生成质量的关键指标。为了充分捕捉工艺参数的关键特征并确保f-CaO标签值数据增强的对抗训练稳定性及快速收敛, 本研究所构建的数据增强在WGAN-GP框架下进行设计, 如图2所示。编码器采用融合了CNN与GRU的混合结构, 首先通过多层卷积及LeakyReLU激活提取多尺度特征,

增强对工艺参数耦合关系的非线性表征; 随后, 将卷积提取的参数特征输入GRU模块, 利用门控机制捕捉长时序依赖并动态更新工艺特征隐藏状态。最后, 通过全连接层映射至潜在特征空间, 实现对高维工艺参数的低维表达。生成器网络则接收编码器的潜在特征空间向量, 通过多级反卷积操作逐层恢复工艺参数数据的空间分辨率, 从而在不同尺度上捕捉工艺参数中的细节并重构目标分布, 生成与实际检测得到高度相似的f-CaO标签数据。判别器网络由多层卷积和线性变换构成, 利用全连接层和卷积层提取工艺参数的局部时空特征与f-CaO标签数据的对应关系, 以区分生成f-CaO标签数据和真实f-CaO标签数据。

通过观察训练过程中判别器的损失函数D_Loss和生成器损失G_Loss的变化趋势, 可以有效判断基于WGAN-GP框架的模型收敛状态。图3为模型的损失函数曲线。训练初期, 生成器无法有效捕获数据中的时空特征, 判别器对于真实数据与生成数据难以进行有效区分, 两者不断博弈, 曲线波动较大且存在突变现象。随着训练的进行, D_loss和G_loss的数值逐渐减小, 当迭代次数到10 000次左右时损失值收敛至0附近并保持稳定。这表明生成器和判别器之间达成了平衡, 判别器无法有效区分真实样本和生成样本, 生成器也无法进一步提高生成样本的质量。在这种情况下, 模型已充分学习到了工艺参数中潜在的分布特征, 生成的样本具有较高的质量和多样性。利用生成器生成的样本进行性能预测, 即可获得更准确和可靠的结果。

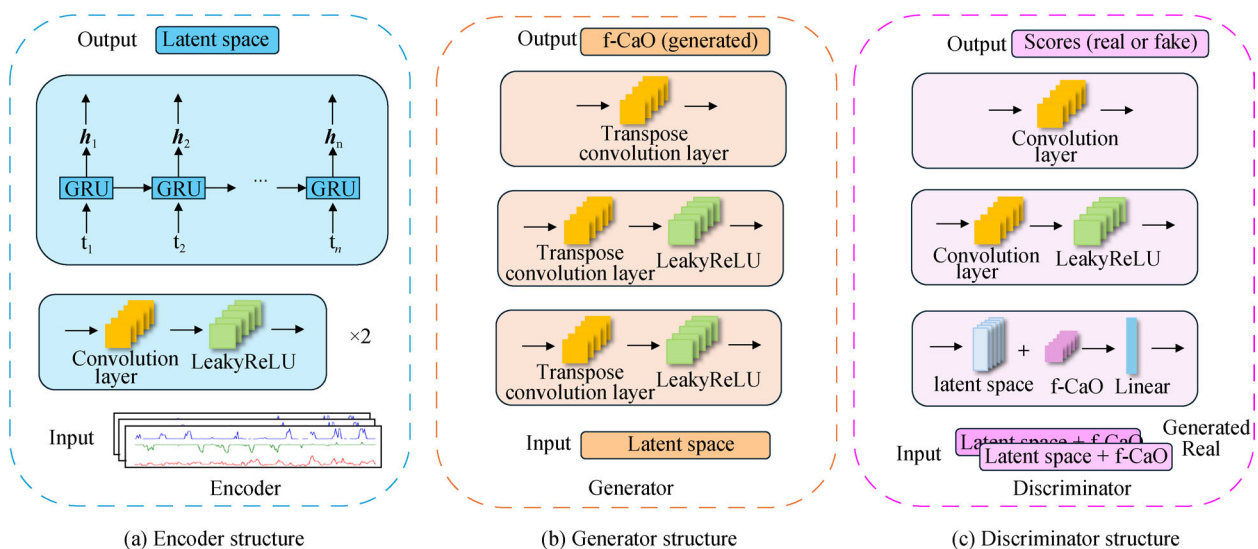


图2 GAN网络结构图
Fig. 2 GAN structure diagram

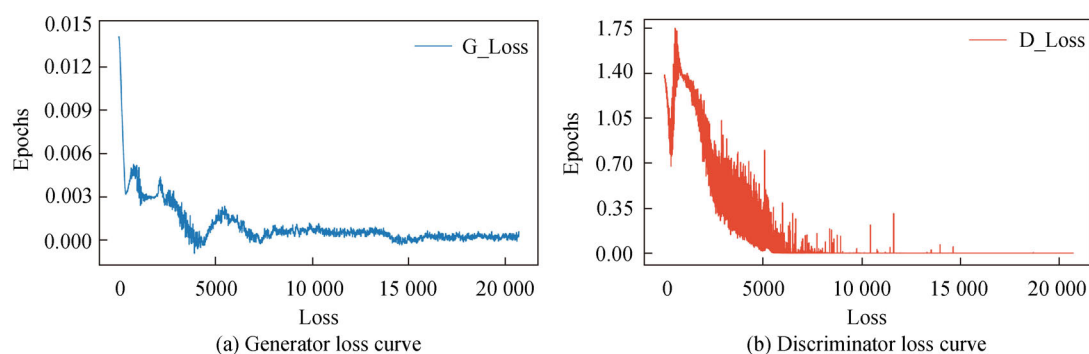


图 3 SSGAN 模型损失函数曲线

Fig. 3 Loss function curve of SSGAN model

2.2 数据增强结果分析

为了评估所提出的 SSGAN 在数据增强方面的有效性, 将生成器生成的标签值与真实标签数据进行了对比。图 4a 和图 4b 分别表示生成器在训练集和测试集上生成标签值的分布情况, 图 5 则为标签值的概率分布曲线, 其中蓝色柱状表示生成标签数据, 橙色柱状表示真实标签数据。

由图 4 和图 5 可见, 在训练集上, 整体模型生成的标签值分布与真实标签值分布在大部分区间内较为接近, 尤其是 0~1 的区间内, 两者的主峰值位置和分布趋势较为一致, 表明模型较好地捕捉到了数据的主要特征。相比之下, 验证集上生成标签值

的分布与真实分布呈现出更为分散的特征, 这在一定程度上反映了模型的泛化能力及数据集本身的差异性。从整体趋势来看, 真实标签值在 0.5 左右出现显著峰值, 生成标签值的主峰与之相对接近, 但在 1.0~2.0 区间仍呈现出一定程度的偏差。综合来看, 所提出的 SSGAN 生成的标签值分布与真实标签值分布大体呈现出相似的分布趋势, 模型在主要标签范围内具有较强的学习与拟合能力, 生成的数据具有可靠性。

在此基础上, 本研究进一步引入滑动时间窗口策略, 对缺失标签的工艺参数矩阵进行标签值生成。通过在时间序列维度上对数据进行动态切片和窗口

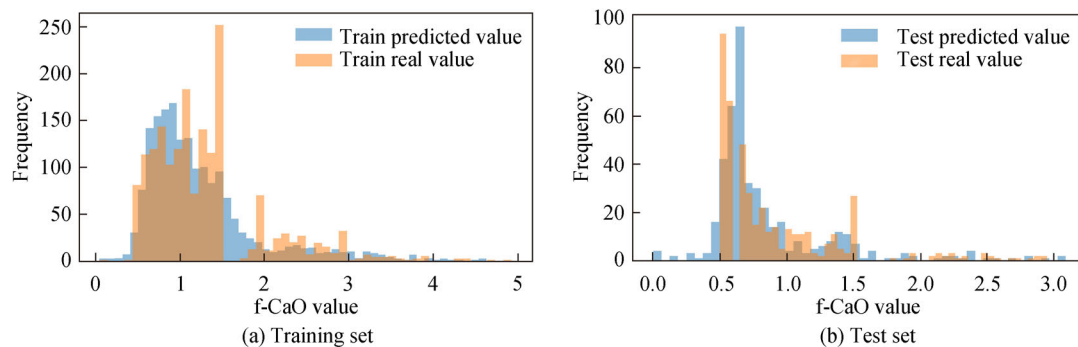


图 4 标签值分布直方图

Fig. 4 Distribution histogram of label values

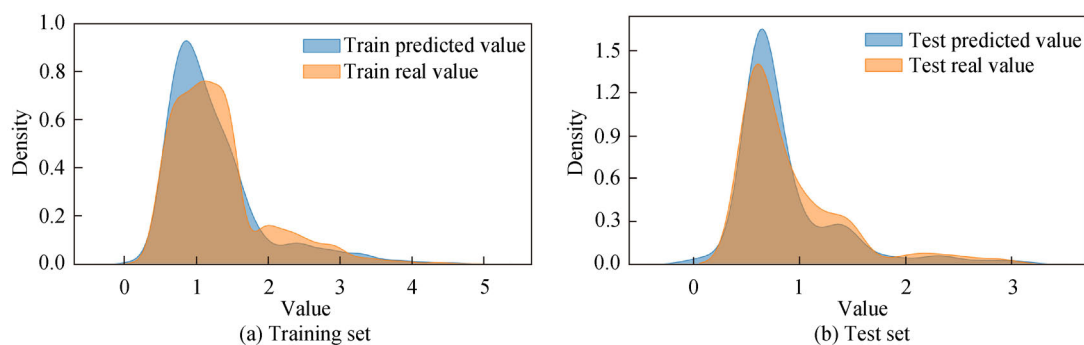


图 5 标签值概率分布图

Fig. 5 Probability distribution of label values

滑动,能够在更细粒度的时刻为无标签的工艺参数分配相应的标签估计值。图6给出了生成标签值与真实标签值的概率分布对比直方图,其反映了此生成过程所得到的标签分布情况及其与真实标签的一致性。由图中可见,生成标签的概率分布大体呈单峰分布,峰值主要集中在0~1的区间,且随着标签值的增大,分布逐渐稀疏,整体呈现一定程度的右偏趋势。相比之下,真实标签在相同区间内也呈现类似的分布特征,峰值位置和整体趋势具有较高的一致性。这与实际生产过程中,工艺控制较为稳定,异常样本量相对较少的情况相符合。上述结果表明通过滑动时间窗口与SSGAN相结合的方法,能够较好地捕捉工艺参数与标签之间的潜在关联特征,实现对缺失标签的合理估计,为后续的预测建模提供了更为丰富和完整的数据基础。

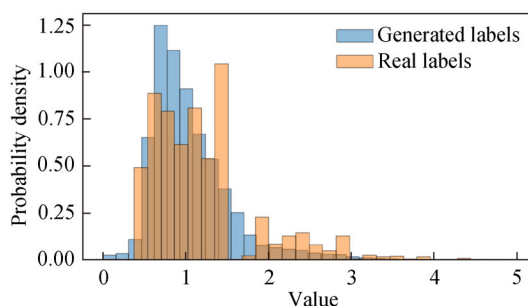


图6 生成与真实标签值的概率分布对比直方图

Fig. 6 Comparison of probability density distributions between generated and real labels

2.3 预测结果分析

为了进一步证明所提出的SSGAN数据增强模型的有效性,本研究在对缺失标签值进行生成的基础上,评估了此策略对于下游预测任务的影响。具体而言,将生成后的数据与原始数据进行融合,以形成更加完备、时序连续性更强的训练集,并在此基础上对GRU模型进行重新训练以预测水泥熟料中的f-CaO含量。图7以及表2分别表明在采用SSGSN数据增强与不采用2种策略下,GRU模型在测试集上的预测结果以及MAE、MSE、RMSE的指标对比。

由表2可知,在引入SSGAN进行数据增强后,GRU模型的MAE、MSE、RMSE均有所下降。这说明通过补充生成的标签值数据,模型对测试集目标值的预测偏差显著减少。同时,由于MSE和RMSE相较于MAE对于预测中的离群点更为敏感,这两项指标的同步降低表明模型对异常值或极端样本

的预测能力也得到了一定程度的提升,整体预测更为稳健。

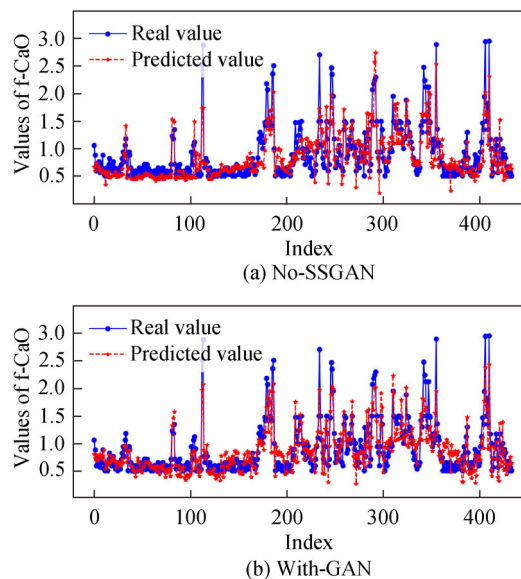


图7 测试集上数据增强处理预测结果

Fig. 7 Data enhancement processes predictive results on test set

表2 测试集上数据增强处理结果

Table 2 Data enhancement processing results on test set

Method	MAE	MSE	RMSE
With-SSGAN	0.19	0.06	0.24
No-SSGAN	0.24	0.08	0.29

SSGAN is semi-supervised generative adversarial network; MAE is mean absolute error; MSE is mean square error; RMSE is root mean square error.

可见,通过在缺失标签的时间段生成合理的标签估计值,可以获得更多的可用训练样本,从而增加样本的数量与标签分布的多样性。这不仅能够缓解模型在小样本场景下可能面临的过拟合风险,也确保了数据与标签在时序上的一致性与完整性,使得模型在训练阶段能够更加充分地学习并捕捉关键的时序依赖特征。

2.4 消融实验

为了进一步探讨SSGAN数据增强模型中模型输入方式及编码器等关键模块对下游预测任务精准度的影响,来研究设计了一系列消融实验,并采用下游预测任务的预测误差作为评价指标进行量化比较。分析实验从以下3个方面进行: Random Noise 指在传统GAN框架下,采用随机噪声作为生成器的条件输入; No Encoder 则取消了编码器的设计,通过简单线性映射直接将原始工艺参数作为

生成器网络的输入; No GRU 在编码器中去除了 GRU 模块, 仅采用 CNN 进行特征提取。实验结果如表 3 所示。

表 3 消融实验结果

Table 3 Ablation results of SSGAN model

Method	MAE	MSE	RMSE
SSGAN	0.19	0.06	0.24
Random noise	0.55	0.65	0.81
No encoder	0.46	0.43	0.66
No GRU	0.40	0.33	0.57

GRU is gate recurrent unit.

由表 3 可以看出, 完整的 SSGAN 模型在 MAE、MSE、RMSE 3 个指标上均取得最低值, 表明了该结构在捕捉工艺参数中时序依赖性及内在模式方面的优势, 从而为下游预测任务提供了高质量的数据支撑。

当完全使用随机噪声替代原始工艺参数输入时, 预测性能显著下降, 表明原始工艺参数中蕴含的时序信息对于生成合理标签数据至关重要, 缺失有效输入信息会导致 f-CaO 标签数据生成质量急剧下降, 进而影响下游任务的预测准确性。

当取消编码器模块, 仅通过简单线性映射处理原始工艺参数数据时, 模型的 MAE、MSE、RMSE 分别为 0.46、0.43、0.66, 虽然性能较随机噪声略有提升, 但与完整模型相比仍存在明显劣势。这表明, 编码器结构在深层特征提取以及时间序列关系建模方面具有不可替代的作用, 简单线性映射难以充分捕捉原始工艺参数数据中的细微特征。其在保证对生产数据进行充分表征的同时, 显著降低了输入空间的复杂度, 为后续生成器和判别器的建模提供了更具针对性的输入特征。

此外, 当移除 GRU 模块, 仅保留 CNN 进行特征提取时, 其预测性能介于随机噪声和无编码器方案之间, 但仍低于完整 SSGAN 模型。这说明 GRU 模块在捕捉工艺参数特征数据中的长程时序依赖和动态变化特征方面起到了关键作用, 能够更准确地提取工艺参数中的隐含信息, 从而生成更为合理的 f-CaO 标签数据。

图 8 给出了消融实验的误差雷达图, 由图可以看出, 完整的 SSGAN 数据增强模型在下游预测任务中表现最佳, 其各项预测误差指标均显著低于其它消融方案。这再一次证明了 SSGAN 中各模块之间的协同作用对生成高质量数据和提升预测模型性能的重要性。

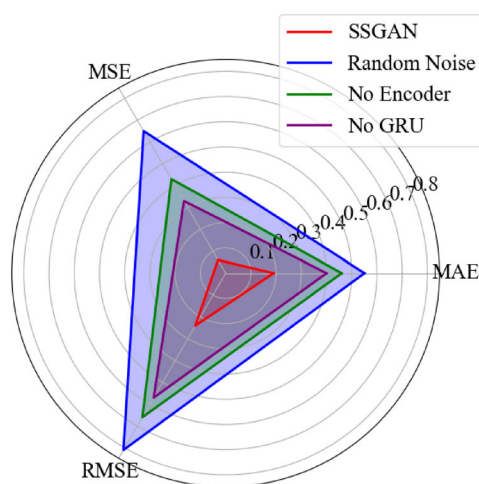


图 8 SSGAN 模型消融实验误差雷达图

Fig. 8 Radar plot of SSGAN model ablation experimental error

3 结论

针对工艺参数与 f-CaO 含量数据采样周期不一致带来的标签值缺失问题, 提出了一种基于 SSGAN 的数据增强方法。通过编码器提取高维特征, 生成器学习真实标签的潜在分布, 构建出更具多样性和代表性的训练数据集, 并结合 GRU 模型进行消融实验, 评估了各关键模块的有效性。主要得到以下结论:

1) 引入 SSGAN 方法数据增强策略后, 有效扩建了训练数据集, 其在主要标签范围内具有较强的学习与拟合能力, 生成的标签值分布与真实标签值分布大体呈现出相似分布趋势, 验证了生成数据的可靠性, 为后续的预测建模提供了更为丰富和完整的数据基础。

2) SSGAN 方法的应用有效提升了下游 f-CaO 含量预测任务的模型精准度, 测试集上的 MAE 由 0.24 下降至 0.19, MSE 由 0.08 降低至 0.06, RMSE 则从 0.29 降至 0.24。

3) 通过消融实验对比表明, 随机噪声、无编码器以及无 GRU 方案的预测误差显著增大, 表明工艺参数中蕴含的时序信息及其深层特征在生成高质量标签数据中的关键作用, 其中编码器及其内部 GRU 模块在捕捉长程时序依赖性和动态变化特征方面尤为重要。

参考文献:

- [1] KURDOWSKI W, THIEL A. On the role of free calcium oxide in expansive cements[J]. Cem Concr Res, 1981, 11(1): 29-40.
- [2] Nawy E G. Concrete Construction Engineering Handbook[M]. 2nd ed.

- Boca Raton: CRC Press, 2008.
- [3] YAO L, JIANG X Y, HUANG G P, et al. Virtual sensing f-CaO content of cement clinker based on incremental deep dynamic features extracting and transferring model[J]. *IEEE Trans Instrum Meas*, 2020, 70: 2500610.
- [4] XU T, ZOU Y C, WANG R Z, et al. Novel f-CaO soft sensor for cement clinker based on integrated model of dual-parallel structure[J]. *Rev Sci Instrum*, 2024, 95(5): 055005.
- [5] 陈亚明. 基于神经网络模拟预分解窑煅烧的研究[D]. 武汉: 武汉理工大学, 2003.
- CHEN Yaming. Study on calcination in precalcining kiln based on Artificial Neural[D]. Wuhan: Wuhan University of Technology, 2003.
- [6] WAN C H, LI J N, FENG L H. Research on RPLS soft-measuring project realization of cement clinker f-CaO[C]//2016 Chinese Control and Decision Conference (CCDC). Yinchuan, China. IEEE, 2016: 4755–4759.
- [7] Wang X, Sun X, Wang Z, et al. Measurement method research for cement f-CaO based on local PSO-LSSVM[J]. *Control Engineering of China*, 2014, 21(6): 807–811.
- [8] HAO X C, LIU L, HUANG G L, et al. R-WGAN-based multitimescale enhancement method for predicting f-CaO cement clinker[J]. *IEEE Trans Instrum Meas*, 2021, 71: 2501310.
- [9] YAN F, ZHANG X M, YANG C J, et al. Data-driven modelling methods in sintering process: Current research status and perspectives[J]. *Can J Chem Eng*, 2023, 101(8): 4506–4522.
- [10] ZHOU X Y, LIU H, CHEN F G, et al. Model-level weight update domain adaptive dynamic CNN soft sensor for free calcium ion concentration in cement clinker[J]. *Chemom Intell Lab Syst*, 2024, 248: 105106.
- [11] ZHAO Y T, DING B C, ZHANG Y L, et al. Online cement clinker quality monitoring: A soft sensor model based on multivariate time series analysis and CNN[J]. *ISA Trans*, 2021, 117: 180–195.
- [12] BAE J, LEE M, KIM S B. Safe semi-supervised learning using a Bayesian neural network[J]. *Inf Sci*, 2022, 612: 453–464.
- [13] Goodfellow I, Pouget-Abadie J, Mirza M, et al. Generative Adversarial Nets[C]//Advances in Neural Information Processing Systems: Vol. 27. Curran Associates, Inc., 2014.
- [14] CRESWELL A, WHITE T, DUMOULIN V, et al. Generative adversarial networks: An overview[J]. *IEEE Signal Process Mag*, 2018, 35(1): 53–65.
- [15] TRAN N T, TRAN V H, NGUYEN N B, et al. On data augmentation for GAN training[J]. *IEEE Trans Image Process*, 2021, 30: 1882–1897.
- [16] YOON J, JORDON J, SCHAAR M. GAIN: Missing Data Imputation using Generative Adversarial Nets[C]//Proceedings of the 35th International Conference on Machine Learning. PMLR, 2018: 5689–5698.
- [17] RADFORD A, METZ L, CHINTALA S. Unsupervised representation learning with deep convolutional generative adversarial networks[J]. *Computer ence*, 2015, 19: 11758569.
- [18] Karras T, Laine S, Aila T. A Style-Based Generator Architecture for Generative Adversarial Networks[J]. *IEEE Trans Pattern Anal Mach Intell*, 2021, 43(12): 4217–4228.
- [19] ZHU J Y, PARK T, ISOLA P, et al. Unpaired Image-To-Image Translation Using Cycle-Consistent Adversarial Networks[C]//Proceedings of the IEEE International Conference on Computer Vision. 2017: 2223–2232.
- [20] 沈威, 黄文熙, 等. 水泥工艺学[M]. 武汉: 武汉工业大学出版社, 1991.
- [21] Li W, Wang D, Chai T. Multisource data ensemble modeling for clinker free lime content estimate in rotary kiln sintering processes[J]. *IEEE T Syst Man Cy-S*, 2015, 45(2): 303–314.
- [22] Zhang Y, Liu J, Dang H, et al. Improving cement production process with data-augmented sequence to sequence-Wasserstein generative adversarial networks model for accurate prediction of f-CaO[J]. *Rev Sci Instrum*, 2023, 94(10): 105111.

A Semi-Supervised Generative Adversarial Network-Based Data Augmentation Method for Cement Clinker f-CaO and Its Effectiveness Evaluation

YAO Xinyue, REN Xuehong, YE Jiayuan, ZHAI Munan, CAO Fuli, ZHANG Wensheng, ZHANG Hongtao
(China Building Materials Academy, Beijing 100024, Chian)

Extended Abstract

Introduction The free lime (f-CaO) content in cement clinker is a key indicator for evaluating calcination quality and process stability, directly influencing cement soundness and mechanical performance. Efficient and accurate real-time prediction of f-CaO content is essential for achieving stable cement quality control. In actual production, process parameter data are sampled at approximately 1-second intervals, while f-CaO content is typically measured once per hour, resulting in a severe scarcity of labeled data. This asynchronous sampling and data heterogeneity constrain both the scale and diversity of data available for supervised learning models, thereby reducing prediction accuracy and generalization capability. To address this challenge, this study proposes a method based on a semi-supervised generative adversarial network (SSGAN) to enhance the diversity and representativeness of the training dataset. Through evaluating its impact on the downstream f-CaO prediction task and conducting ablation studies, the research explores the feasibility of applying GANs to address the issue of asynchronous sampling between process parameters and offline measurement data.

Methods This study proposes a data augmentation model based on a semi-supervised generative adversarial network (SSGAN), which generates virtual data resembling the actual data distribution to expand the training dataset and effectively enhance the model's learning capability under conditions of limited labeled data. The model consists of three main components: an encoder, a generator

network, and a discriminator network. The encoder first extracts features from a small amount of process parameter data with real labels, which are then used as conditional inputs. The generator learns the latent features of the true label distribution through adversarial training and generates corresponding label values based on unlabeled process parameter data. The discriminator distinguishes between real and generated labels, guiding the generator to continuously optimize its generation strategy. As adversarial training progresses, the generator produces synthetic label data that closely resemble actual production data, while the discriminator continuously improves its understanding of the true data distribution and label characteristics.

Results and discussion When the SSGAN model is trained to approximately 10 000 iterations, both the discriminator loss and the generator loss converge to near zero and remain stable, indicating that a balance has been achieved between the generator and discriminator, and the model has effectively learned the latent distribution features within the process parameters. Further analysis of the distribution of generated label values shows that they exhibit a distribution trend generally similar to that of the real labels. The model demonstrates strong learning and fitting capability within the primary range of label values, and the generated data are reliable. After introducing SSGAN for data augmentation, the MAE, MSE, and RMSE of the downstream GRU (Gate Recurrent Unit) model for f-CaO content prediction decreased from 0.24, 0.08, and 0.29 to 0.19, 0.06, and 0.24, respectively, indicating an improvement in the model's ability to predict outliers and extreme samples. Ablation studies revealed that removing the encoder, the GRU module, or introducing random noise led to significantly higher prediction errors, highlighting the critical role of temporal information and deep features embedded in the process parameters for generating high-quality label data. In particular, the encoder and its internal GRU module are essential for capturing long-range temporal dependencies and dynamic variation patterns.

Conclusions To address the issue of label scarcity caused by the mismatch in sampling intervals between process parameters and f-CaO content measurements, this study proposes a data augmentation method based on SSGAN. By using an encoder to extract high-dimensional features and a generator to learn the latent distribution of true labels, the approach constructs a more diverse and representative training dataset. Ablation experiments were conducted to evaluate the effectiveness of each key module. The main conclusions are as follows: The introduction of the SSGAN-based data augmentation strategy effectively expanded the training dataset. The distribution of the generated label values closely matched that of the true labels, verifying the reliability of the synthetic data. The application of the SSGAN method significantly improved the accuracy of the downstream f-CaO content prediction task, with MAE on the test set decreasing from 0.24 to 0.19, MSE from 0.08 to 0.06, and RMSE from 0.29 to 0.24. Ablation experiments further revealed that introducing random noise, removing the encoder, or omitting the GRU module led to significantly increased prediction errors. This highlights the critical role of temporal information and deep features embedded in process parameters for generating high-quality label data. In particular, the encoder and its internal GRU module are essential for capturing long-range temporal dependencies and dynamic variation patterns.

Keywords free lime; semi-supervised generative adversarial network; time series; data enhancement; regression forecasting; ablation experiment