基于多模态时序对比生成网络的数据增强算法

商 柔[1,2,3], 董宏丽[2,3†], 王 闯[2,3], 周国强[1], 管 闯[2,3], 闫天红[1]

(1. 东北石油大学 三亚海洋油气研究院, 海南 三亚 572025; 2. 东北石油大学 人工智能能源研究院, 黑龙江 大庆 163318;
3. 黑龙江省网络化与智能控制重点实验室, 黑龙江 大庆 163318)

摘要: 针对工业故障诊断中的小样本和类不平衡问题, 本文提出一种基于马尔可夫链的多模态时序对比生成模型(TCGN). 首先, 为了提升合成数据时间结构的真实性, 设计了一种时序趋势一致化损失(TTC), 以提升真实数据与合成数据之间时间演化规律的相似度. 随后, 为了在增强数据集中形成有效且正确的决策边界, 提出了一种类意识对比损失(CAC), 以对齐真实数据与合成数据的类条件分布. 此外, 为了更好地维持不同学习任务之间的动态平衡, 引入了一种基于马尔可夫链的多模态切换策略, 以实现TCGN算法在生成、刻画、探索、收敛4个模态之间的自适应切换优化. 最后, 将所提出的TCGN算法应用于管道故障诊断. 实验结果表明TCGN算法在视觉评估和量化指标方面均优于一些先进的生成算法, 显著提高了故障诊断准确率.

关键词: 管道故障诊断; 类别不平衡; 时间序列; 数据增强; 马尔可夫链; 多任务学习

引用格式: 商柔, 董宏丽, 王闯, 等. 基于多模态时序对比生成网络的数据增强算法. 控制理论与应用, 2025, 42(4): 805 – 815

DOI: 10.7641/CTA.2023.30066

Multimodal time-series contrastive generative network-based data augmentation algorithm

SHANG Rou[1,2,3], DONG Hong-li[2,3†], WANG Chuang[2,3],
ZHOU Guo-qiang[1], GUAN Chuang[2,3], YAN Tian-hong[1]

(1. Sanya Offshore Oil & Gas Research Institute, Northeast Petroleum University, Sanya Hainan 572025, China;
2. Artificial Intelligence Energy Research Institute, Northeast Petroleum University, Daqing Heilongjiang 163318, China;
3. Heilongjiang Provincial Key Laboratory of Networking and Intelligent Control,
Daqing Heilongjiang 163318, China)

Abstract: In this paper, a Markov chain-based multimodal time-series contrastive generative network (TCGN) is proposed to tackle the issues of small sample and class imbalance for industrial fault diagnosis. Firstly, a time-series trend consistency loss (TTC) is designed to enhance the similarity of the time-evolving properties between the real and synthetic data, which helps to improve the reality of the synthetic temporal structure. Subsequently, a class-aware contrastive loss (CAC) is proposed to align the class-conditional distributions between the real and synthetic datasets, which facilitates the formation of effective and proper decision boundaries. Furthermore, a Markov chain-based multimodal switching strategy is introduced in this paper, which enables the TCGN algorithm to perform adaptive switching optimization between the four modes of generation, depiction, exploration, and convergence, thus better maintaining the dynamic balance of different tasks. Finally, the proposed TCGN algorithm is applied to pipeline fault diagnosis. Experimental results show that the TCGN algorithm outperforms some state-of-the-art generation algorithms in terms of visual evaluation and quantitative metrics, and significantly improves fault diagnosis accuracy.

Key words: pipeline fault diagnosis; class imbalance; time series; data augmentation; Markov chain; multi-task learning

Citation: SHANG Rou, DONG Hongli, WANG Chuang, et al. Multimodal time-series contrastive generative network-based data augmentation algorithm. *Control Theory & Applications*, 2025, 42(4): 805 – 815

收稿日期: 2023−02−15; 录用日期: 2023−12−12.

†通信作者. E-mail: shiningdhl@vip.126.com; Tel.: +86 459-6503373.

本文责任编委: 岳东.

国家自然科学基金项目(U21A2019, 61873058, 61933007), 海南省科技专项项目(ZDYF2022SHFZ105)资助.

Supported by the National Natural Science Foundation of China (U21A2019, 61873058, 61933007) and the Hainan Province Science and Technology Special Fund (ZDYF2022SHFZ105).

1 引言

管道作为一种有效的长距离运输载体已广泛应用于石油与天然气工业, 是我国实施能源战略的重要基础设施. 在实际生产中, 为满足运输需求和节省土地资源, 管道通常埋藏于地下, 土壤腐蚀和自然灾害等因素不可避免会导致管道发生穿孔、破损等问题. 若未能及时发现并修复, 会引发大量能源泄漏, 轻则缩短管道使用寿命、增加经济成本, 重则造成环境污染、人员伤亡. 因此, 研究如何准确检测和识别管道健康状态, 对减少安全事故的发生、提高能源运输的安全性与稳定性, 具有重要的现实意义[1–2].

近些年, 随着深度学习技术的兴起, 各领域学者开始关注和研究基于数据驱动的智能诊断方法[3–4]. 然而, 这些先进算法在应用于实际工业场景时并没有展现出理想的性能, 原因在于: 1) 智能诊断方法有效的前提是训练集中具有充足的故障样本, 但管道无法持续运行在故障状态, 因此收集的监测数据中故障样本严重稀缺且类别不全; 2) 高质量的标签是训练有监督分类模型的必要条件, 但人工标注需要花费大量的时间成本和高昂的经济成本且无法保证时效性, 导致数据集内的可用标记样本数量不足. 从上述讨论中可以发现, 实际场景中可收集到的数据集通常呈现小样本, 类分布偏斜特性. 现有的先进算法在处理样本量充足、类均衡、决策边界清晰的数据集时往往可以取得较好的分类性能. 然而当面对小样本且类不平衡数据集时, 模型会倾向于忽略小类故障样本来提高对于大多数样本的拟合能力, 从而导致准确率降低, 误报率与漏报率升高. 因此, 研究如何有效地解决小样本和类不平衡问题具有十分重要的指导意义和实际应用价值.

近年来, 生成对抗网络 (generative adversarial net, GAN)[5]在学习复杂高维的真实世界数据分布方面表现出了极其优越的性能, 为解决故障诊断领域的小样本和类不平衡问题提供了一个新视角, 引发了学者们的研究热潮. 通过生成器与判别器的对抗学习, GAN可以捕捉真实数据的分布特征, 生成具有与之相似分布的合成数据, 从而有效解决多类不平衡问题, 提升诊断模型性能. 然而, 传统GAN算法在处理时间序列方面仍有许多困难, 无法有效地对高维向量进行特征建模. 为解决上述问题, Esteban等人[6]提出引入递归神经网络(recursive neural network, RNN)到生成器与判别器中, 以产生现实的实值多维时间序列. Wang等人[7]引入长短时记忆网络 (long short-term memory, LSTM)改进原始GAN算法的生成器, 使其适用于学习长期依赖信息, 提高合成数据质量. 尽管先进GAN算法已经尝试利用时序网络改进传统GAN结构, 但RNN和LSTM的训练很难达到全局最优, 导致其不能有效地合成时间结构. 其次, 管道运行工况复杂多变, 采集的数据特征也不尽相同. 例如, 同类别管道数据可能波形相似, 但幅值范围相差较大, 或是不同类别管道数据可能波形差异明显, 但幅值范围相差较小. 在这种情况下, 仅提高全局相似性而不考虑局部相关性, 往往会使相同类别的合成数据与真实数据不匹配. 综上所述, 无论是在实例层面还是类别层面, 现有的GAN算法均难以保证其在管道数据增强中可以展现优越的生成性能.

为了提升合成数据的质量, 缓解小样本和类不平衡问题带来的负面影响, 本文提出了一种基于马尔可夫链的多模态时序对比生成模型(time-series contrastive generative network, TCGN). 该模型旨在学习整体分布特征的同时, 既能考虑合成数据时间趋势的真实性, 又可以兼顾类别分布的统一性, 以此提升故障诊断性能. 本文的主要贡献总结如下: 1) 在实例层面, 提出了一种时序趋势一致化损失(time-series trend consistency loss, TTC), 通过学习真实数据的时间特征, 提高合成时间结构的可靠性; 2) 在类别层面, 提出了一种类意识对比损失 (class-aware contrastive loss, CAC), 通过对齐相应健康状态下的合成分布与真实分布, 提升类内紧凑性和类间可分离性, 进而在增强数据集中形成有效且清晰的决策边界; 3) 为了维持不同学习任务之间的动态平衡, 本文将TCGN的训练过程划分为生成、刻画、探索、收敛4个模态, 并引入马尔可夫链实现了模态之间的自适应切换优化; 4) 实验结果表明, TCGN在定性特征可视化与定量指标评估中明显优于一些先进算法, 成功解决了小样本和类不平衡问题, 提升了管道故障诊断准确率.

2 基于马尔可夫链的多模态时序对比生成模型

围绕管道数据增强任务, 本文提出了一种同时学习真实数据时间结构和局部分布类别特征的时序对比生成模型. 其中, 为了维持不同学习任务之间的动态平衡, 设计了一种基于马尔可夫链的多模态切换策略, 有效提高了合成数据的质量. TCGN算法的整体结构如图1所示. 下面对TCGN算法的原理进行详细说明.

2.1 时序趋势一致化损失

考虑到GAN算法在非时间序列数据上的巨大成功, 许多研究人员尝试将现有的非时间序列数据的工作直接拓展到时间序列数据的场景中. 然而, 由于缺乏对时间动态特性的认知能力[8], 这些算法在生成时间序列数据方面的性能有限, 无法满足实际应用需求. 为解决这一问题, 本文提出了一种新的TTC损失, 目的是提升生成模型合成时间关联结构的能力. TTC损失的过程可以表述为: 1) 根据振幅变化趋势将数据分成不同长度的子序列; 2) 依据分割后序列的倾斜程度

将子序列划分为不同的模式; 3) 计算重构后的真实数据与合成数据的相似性. 具体计算过程如下所示.

假设长度为 L 的管道时序数据可以表示为 $X=\{x_1,x_2,\cdots,x_L\}$, 则分段后的时间序列可以表征为 $\hat{X}=\{\hat{m}_1,\hat{m}_2,\cdots,\hat{m}_S\}$, $\hat{m}$是子序列模式, S代表一个时间序列样本被分割后的区段数. 因此, 第j个真实数据与合成数据之间的时序趋势相似性可描述为

$$\mathcal{L}_{\text{ttc}}^{i}=\sum_{s=1}^{S}(w_{i,s}\times|\hat{m}_{i,s}^{\text{g}}-\hat{m}_{i,s}^{\text{r}}|\times|A_{i,s}^{\text{g}}-A_{i,s}^{\text{r}}|),\tag{1}$$

式中: $w_{i,s}=l_{i,s}/L$是权重系数, $l_{i,s}$代表第 i 个样本、第s 区间段时间序列的长度; g和r分别代表合成数据和真实数据; $\hat{m}_{i,s}$ 代表第i个样本、第s区间段子序列的模式, 具体地说, $\hat{m}_{i,s}=-1$代表下降趋势, $\hat{m}_{i,s}=0$代表平稳趋势, $\hat{m}_{i,s}=1$代表上升趋势; $A_{i,s}=y_{s+l_s}-y_s$是第s个区间段的端点幅值差. 由此, 可以得到总体时序趋势一致化损失为

$$\mathcal{L}_{\text{ttc}}=\frac{1}{N}\sum_{i=1}^{N}\mathcal{L}_{\text{ttc}}^{i},\tag{2}$$

式中N是生成样本数. TTC损失的目的是突出时间序列的趋势变化并弱化由于空间转换或人为选择偏差导致的序列异常点, 从而加强对于时间演化规律的学习能力, 进一步提升合成数据的真实性和模型鲁棒性.

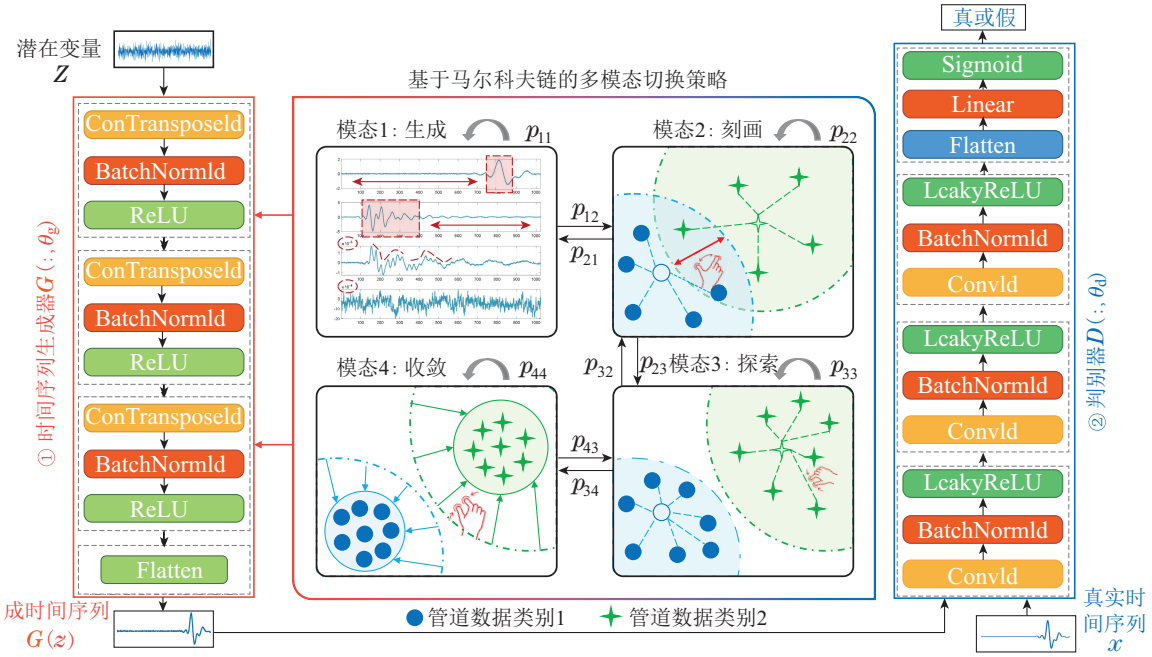


图 1　TCGN算法的网络结构

Fig. 1　Network structure of TCGN algorithm

2.2　类意识对比损失

传统GAN算法主要关注如何通过对抗学习提升真实分布与合成分布之间的全局相似性, 而忽略了对应健康状态下真实子分布和合成子分布之间的关系. 在这样的情况下, 来自不同类别的真实数据与合成数据的判别性结构会被混淆, 致使细粒度信息丢失. 因此, 现有的生成模型难以准确恢复故障数据特征, 导致增强数据集可靠性不高. 考虑到对应子分布的关系, 本文分析了增强数据集的特性: 1) 同一类别中的合成数据应保持一致, 不同类别的合成数据之间应具有明显区分性; 2) 真实数据集和合成数据集中的同类别子分布应保持对齐, 不同类别子分布之间应存在一个清晰且有效的决策边界.

为了提升增强数据集的可判别性, 本文提出了一种基于最大均值差异 (maximum mean discrepancy, MMD)[9]的CAC损失来学习局部分布的判别特征, 迫使合成数据向其相同类别的真实数据聚拢, 同时扩大不同类别合成数据与真实数据之间的距离. 具体实现方式如下所述. 首先, 不同类别合成数据与真实数据的对应分布差异可以表示为

$$\mathcal{L}_{\text{cac}}^{\text{inter}}=\frac{1}{C}\sum_{c=1}^{C}\Big\|\sum_{G(z|y)_j\in P_c}\varphi(G(z\mid y)_j)-\frac{1}{C-1}\sum_{x_j\notin P_c}\varphi(x_j)\Big\|_{\mathcal{H}},\tag{3}$$

式中: $\mathcal{H}$代表再生希尔伯特空间 (reproducing kernel Hilbert space, RKHS); C代表类别总数; y是类别标签; φ代表从样本空间到RKHS的非线性映射. 式(3)计算的是c类别的合成数据与所有非此类别的真实数据之间的差异, 可以称之为类间差异. 通过优化类间差异, TCGN算法可以学习到清晰的决策边界, 从而较好地刻画不同类别分布间的决策边界. 其次, 类内损失的

计算方式如下:

$$\mathcal{L}_{\text{cac}}^{\text{intra}}=\frac{1}{C}\sum_{c=1}^{C}\|\sum_{G(z|y)_j\in P_c}\varphi(G(z\mid y)_j)-\sum_{x_j\in P_c}\varphi(x_j)\|_{\mathcal{H}},\tag{4}$$

式(4)计算的是同类别合成数据与真实数据之间的差异, 可以称为类内差异. 通过缩小类内差异, TCGN算法可以获得更紧凑的样本特征, 从而进一步提升分类准确率.

为了明确合成样本的类别信息, 本文采用条件GAN网络(conditional generative adversarial networks, CGAN)[10]作为TCGN算法的基线. 因此, 时间序列生成器的损失函数为

$$\mathcal{L}_{\text{G}}^{\text{adv}}=\text{E}_{z\sim P_z(z)}[\log(1-D(G(z\mid y)\mid y))].\tag{5}$$

判别器损失函数为

$$\mathcal{L}_{\text{D}}^{\text{adv}}=-\text{E}_{x\sim P_{\text{data}}(x)}[\log D(x\mid y)]-\text{E}_{z\sim P_z(z)}[\log(1-D(G(z\mid y)\mid y))].\tag{6}$$

2.3 基于马尔可夫链的多模态切换策略

遵循大多深度学习算法所采用的优化策略, TCGN中生成器的总体损失函数可以表述为如下形式:

$$\mathcal{L}_{\text{G}}=\mathcal{L}_{\text{G}}^{\text{adv}}+\alpha\mathcal{L}_{\text{ttc}}-\beta\mathcal{L}_{\text{cac}}^{\text{inter}}+\gamma\mathcal{L}_{\text{cac}}^{\text{intra}},\tag{7}$$

式中α, β, γ是平衡超参数. 尽管TTC与CAC可以提高合成数据的质量, 但式(7)所示的优化策略对于TCGN的训练来说并不是最佳方案, 原因如下.

首先, TCGN的学习任务可以概括为以下4点: 1) 提升合成时序结构的真实性; 2) 扩大不同类别合成数据与真实数据之间的差异性; 3) 增强相同类别合成数据与真实数据之间的相关性; 4) 提高合成数据与真实数据之间的全局相似性. 在训练过程中, 个体任务的困难程度可能导致所有任务的训练进度不一致. 因此, 根据式(7)进行优化, 往往会限制某个阶段的任务学习, 导致生成性能退化或停滞. 其次, 引入TTC损失与CAC损失在一定程度上加大了生成模型的优化压力, 增加了模型过拟合的风险. 此外, 多超参数组合寻优很难找到最优解, 导致生成性能不佳. 综上所述, 传统的优化策略难以适配TCGN训练, 可能会降低增强数据集的整体质量.

为了解决这个关键问题, 本文根据上述学习任务将TCGN的生成器分为4种不同的模态, 即生成、刻画、探索和收敛, 并引入马尔可夫链实现4种模态的自适应切换优化, 提升训练稳定性[11].

马尔可夫链的原理如下: 假设$\xi(h)(h\in\mathbb{N}$, $\mathbb{N}$为非负整数集) 是在有限状态空间$\mathcal{S}=\{1,2,\cdots,M\}$中取值的马尔可夫链, 其概率转移矩阵$P=(p_{ij})_{M\times M}$可以通过下式得出:

$$P=\{\xi(h+1)=j\mid\xi(h)=i\}=p_{ij},\tag{8}$$

式中$p_{ij}\geqslant 0$ $(i,j\in\mathcal{S})$代表从i到j的转移概率, 且$\sum\limits_{j=1}^{M}p_{ij}=1$.

基于马尔可夫链的多模态切换策略的原理如下: 根据生成器当前迭代次数下的所属模态来预测下一时刻的生成器模态, 从而完成不同任务之间的切换学习. 为了实现这一过程, 本文引入演化因子E_{f}来判定生成器所属模态, E_{f}的定义如下:

$$E_{\text{f}}(F_h)=\frac{F_h-F_{\min}}{F_{\max}-F_{\min}},\tag{9}$$

式中: F_h代表第h次迭代时合成数据的评估分数; $F_{\max}$和$F_{\min}$分别是h次迭代内评估分数的最大值和最小值. 根据文献[12], 评估分数F_h的计算方式如下:

$$F_h=\text{E}_z[D(G(z\mid y)\mid y)]-\lambda\log\|\nabla_D-\text{E}_x[\log D(x\mid y)]-\text{E}_z\left[\log(1-D(G(z\mid y)\mid y))\right]\|,\tag{10}$$

式中: 第1项是判别分数均值, 代表合成数据的真实性; 第2项是判别器的负对数梯度范数, 表征合成数据的多样性; λ是平衡质量与多样性的超参数.

根据演化因子E_{f}的值, 可以评估生成器的当前模态, 如下所示:

$$\xi(h)=\begin{cases}1,\ 0\leqslant E_{\text{f}}(F_h)<0.25,\\ 2,\ 0.25\leqslant E_{\text{f}}(F_h)<0.5,\\ 3,\ 0.5\leqslant E_{\text{f}}(F_h)<0.75,\\ 4,\ 0.75\leqslant E_{\text{f}}(F_h)\leqslant 1,\end{cases}\tag{11}$$

式中: $\xi(h)=1$代表生成模态; $\xi(h)=2$代表刻画模态; $\xi(h)=3$代表探索模态; $\xi(h)=4$代表收敛模态. 具体细节描述如下:

1) 生成模态$\{\xi(h)=1\}$.

TCGN算法的主要目的是完成从无序潜在变量到时序结构的映射. 因此, 在生成模态, TCGN联合优化了生成器损失与TTC损失, 以提升合成数据时间维度的真实性, 如下所示:

$$\mathcal{L}_1=\mathcal{L}_{\text{G}}^{\text{adv}}+\alpha\mathcal{L}_{\text{ttc}},\tag{12}$$

式中α是平衡生成损失和TTC损失的超参数.

2) 刻画模态$\{\xi(h)=2\}$.

TCGN算法的任务是学习不同类别真实数据和合成数据之间的差异, 促进模型学习决策边界. 在此模态, TCGN联合优化了生成器损失与类间损失, 如下所示:

$$\mathcal{L}_2=\mathcal{L}_{\text{G}}^{\text{adv}}-\beta\mathcal{L}_{\text{cac}}^{\text{inter}},\tag{13}$$

式中β是控制生成损失与类间损失平衡的超参数.

3) 探索模态$\{\xi(h)=3\}$.

TCGN算法的主要任务是加强同类别合成数据与真实数据之间的相互关联, 提升合成序列判别特征的可靠性. 该模态的损失函数描述如下:

$$\mathcal{L}_3=\mathcal{L}_{\mathrm{G}}^{\mathrm{adv}}+\gamma\mathcal{L}_{\mathrm{cac}}^{\mathrm{intra}}, \tag{14}$$

式中γ是调整生成损失和类内损失比例的超参数.

4) 收敛模态$\{\xi(h)=4\}$.

为了加速生成器进入全局最优区域的进程, 在此模态, 本文仅采用了单一生成损失, 加快模型收敛速度. 该模态的损失函数可表述为如下形式:

$$\mathcal{L}_4=\mathcal{L}_{\mathrm{G}}^{\mathrm{adv}}. \tag{15}$$

在训练的开始阶段, 设定生成器为生成模态, 在经过演化因子评估后, 马尔可夫链会以一定的概率预测生成器在下一阶段的所属模态, 此过程的转移概率矩阵可以定义为

$$P=\begin{pmatrix}0.9 & 0.1 & 0 & 0\\ 0.05 & 0.9 & 0.05 & 0\\ 0 & 0.05 & 0.9 & 0.05\\ 0 & 0 & 0.1 & 0.9\end{pmatrix}. \tag{16}$$

3　TCGN生成性能实验结果与分析

3.1　实验设置

本文所使用的训练数据采自于管道仿真平台中的声波传感器. 管道总长度为180 m, 实验过程中, 设置压力为0.5 MPa, 流速为10 $\mathrm{m^3/h}$. 根据阀门开合度, 数据集中包括大泄漏、中泄漏、小泄漏、正常4种健康状态, 总样本量为2 000, 每种健康状况下样本量为500.

TCGN算法由两个模型组成, 即判别器和生成器, 其架构如表1和表2所示.

表 1　判别器模型结构

Table 1　Model architecture of discriminator

编号	网络层	卷积核大小	步长	参数
1	Conv1d	16 × 1	16 × 1	8 389 120
2	BatchNorm1d	—	—	1024
3	LeakyReLU	—	—	—
4	Conv1d	3 × 1	1 × 1	196 736
5	BatchNorm1d	—	—	256
6	LeakyReLU	—	—	—
7	Conv1d	2 × 1	1 × 1	4 112
8	BatchNorm1d	—	—	32
9	LeakyReLU	—	—	—
10	Flatten	—	—	—
11	Linear	—	—	130
12	Sigmoid	—	—	—

在TCGN算法的训练过程中, 总迭代次数为2000; 每次循环中, 判别器训练次数为5, 生成器训练次数为10; 批量大小为100; 训练采用的优化器为Adam, 学习率为1e−4; α, β, γ均为0.01; 根据文献[12], 平衡超参数λ为0.001. 本文中所提及的算法均由PyTorch框架搭建, 在NVIDIA GEFORCE RTX 3090 GPU上训练.

表 2　生成器模型结构

Table 2　Model architecture of generator

编号	网络层	卷积核大小	步长	参数
1	ConvTranspose1d	3 × 1	2 × 1	77 056
2	BatchNorm1d	—	—	512
3	ReLU	—	—	—
4	ConvTranspose1d	3 × 1	2 × 1	393 728
5	BatchNorm1d	—	—	1 024
6	ReLU	—	—	—
7	ConvTranspose1d	3 × 1	2 × 1	1 573 888
8	BatchNorm1d	—	—	2 048
9	ReLU	—	—	—
10	Flatten	—	—	—

3.2　对比算法

为了全面评估TCGN算法的生成性能, 本文选取了7种数据增强方法进行对比研究. 第1类对比算法为传统插值方法, 即合成少数类过采样 (synthetic minority oversampling technique, SMOTE)算法[13]. 第2类对比算法为基于深度学习的方法, 即变分自编码器 (variational auto-encoders, VAE)[14]. 第 3 类对比算法为未考虑时间结构的GAN方法, 包括最小二乘GAN算法(least squares GAN, LSGAN)[15]、基于辅助分类器的GAN算法(auxiliary classifier GAN, ACGAN)[16]、InfoGAN[17]. 第4类对比算法为考虑时间结构的基于GAN的方法, 包括TTS-GAN[18]和MAD-GANs[19].

3.3　定性可视化分析

为了比较生成的样本和相应的真实样本之间的相似性, 本文将生成的管道数据以折线图的方式进行呈现, 如图2所示, 其中第1行为大泄漏, 第2行为中泄漏, 第3行为小泄漏, 第4行为正常状态. 首先, 从图2(a)–(b)中可以看出, SMOTE算法与VAE算法很难恢复真实数据时间维度上的振幅变化, 导致合成序列仅具有大致趋势而不具备有效判别特征. 图2(c)–(e)中的合成数据虽然在变化趋势与幅值方面都与真实数据接近, 但这些算法弱化了管道数据的时间属性, 导致合成数据中包含过多的噪声, 无法用于扩充管道数据集. 此外, 由于TTS-GAN 算法与 MAD-GANs 算法采用了适用于学习时间序列的网络结构, 因此生成的数据较之前的算法特征更加显著, 但仍包含一定的噪声, 如图2(f)–(g)所示. 不同于上述对比算法, TCGN算法很好地学习了样本空间的时间信息与判别特征, 这使得合成管道数据的局部细节与整体结构都更加自然和平滑.

3.4 定量统计指标分析

为了进一步验证合成数据的有效性, 在本小节中, 利用MMD作为量化生成算法性能的一个统计指标. MMD的基本思想在于: 以两个分布之间差距最大的阶矩作为度量两个分布相似性的标准. 一般来说, 该值越小就意味着两个分布越相似.

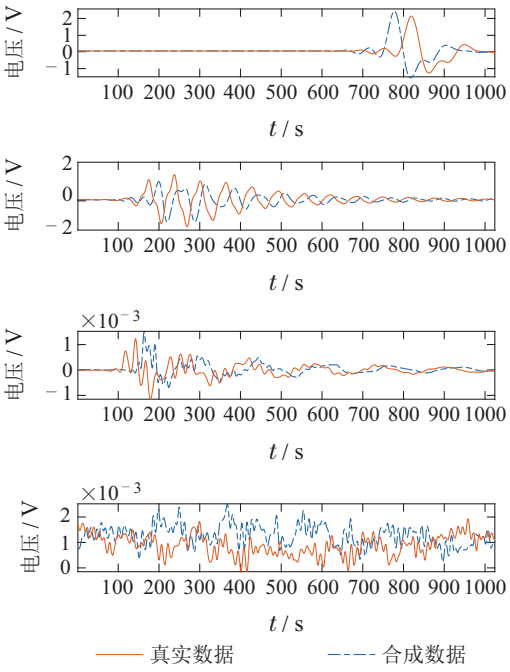
图3展示了不同健康状态下的实验结果. 从柱状图的分布中可以看出, TCGN算法明显优于所有的对比算法, 在4种健康状态中均取得了最佳的统计结果. 具体来说, SMOTE算法和VAE算法的所有健康状态MMD值均超过3.0. 同时, LSGAN, ACGAN, InfoGAN, TTS-GAN的MMD值均高于2.0. 此外, 从图中可以看出, MAD-GANs取得了相对较好的统计结果. 尽管如此, TCGN算法的大泄漏、小泄漏、正常状态的MMD值均在1.5以下, 小于几乎所有算法的MMD值(中泄漏状态与MAD-GANs持平). 通过以上分析, 可以得出结论: TCGN算法生成的管道数据, 其统计特性更接近于真实数据.

4 故障诊断实验结果与分析

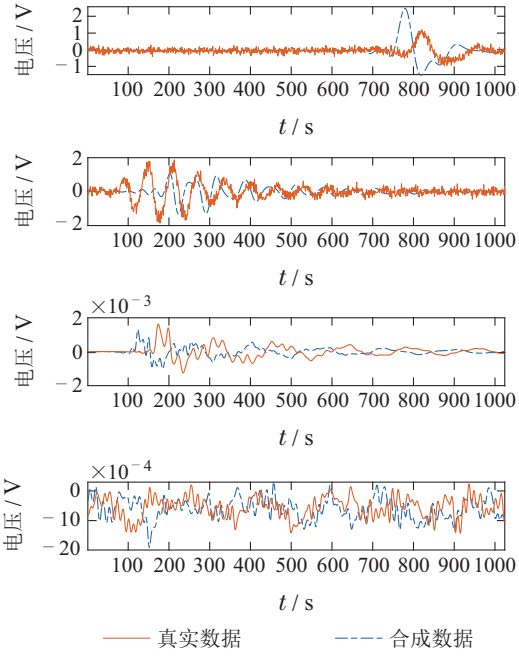
4.1 模型训练与实现细节

管道故障诊断是管道安全运维中十分重要的一环, 对于延续管道寿命和降低运输不稳定性至关重要[20]. 为了进一步证明所提出的针对小样本和类不平衡数据集的TCGN算法的优越性, 本文在不同增强率(augmentation rate, AR)下进行了故障诊断实验. AR的定义如下:

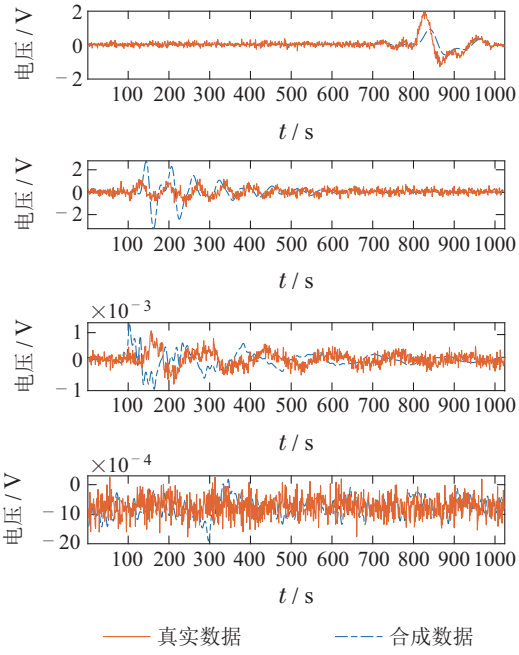
$$\text{增强率} = \frac{\text{正常状态的数量}}{\text{单一故障状态的数量}}. \tag{17}$$



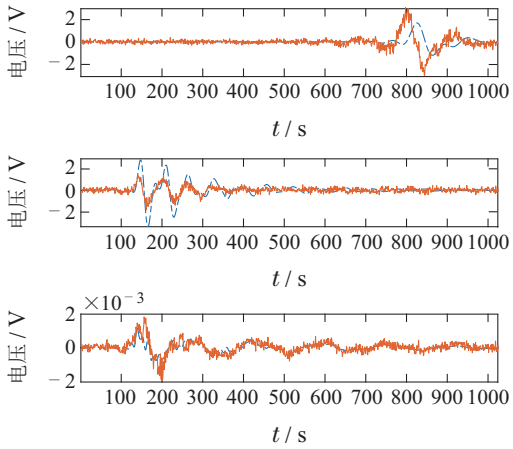
(a) SMOTE

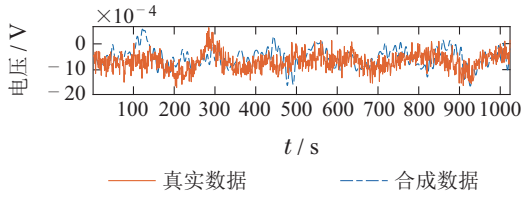


(b) VAE

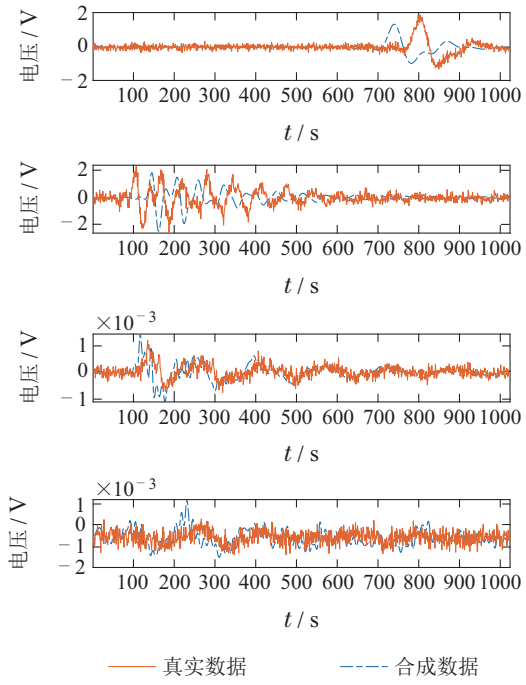


(c) LSGAN

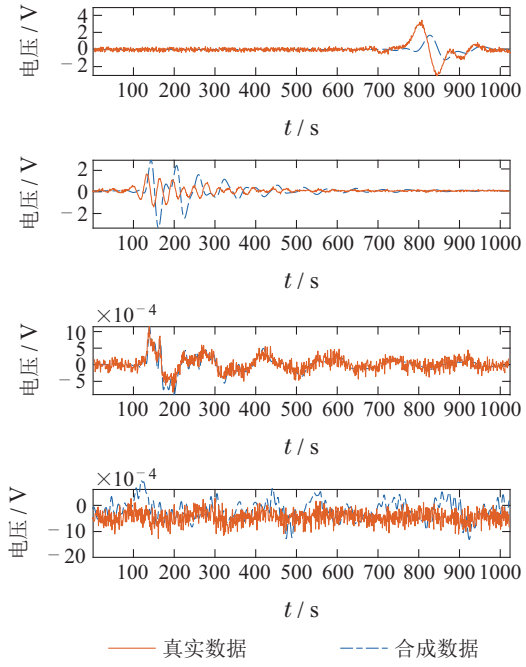




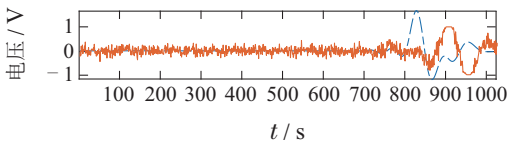
(d) ACGAN

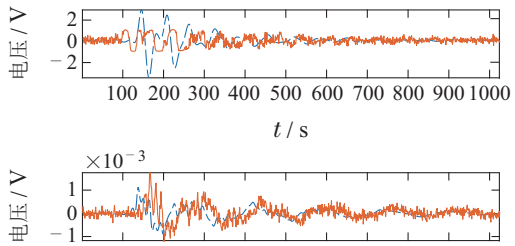


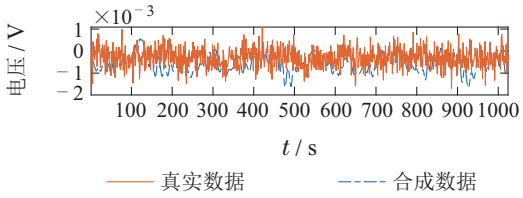
(e) InfoGAN



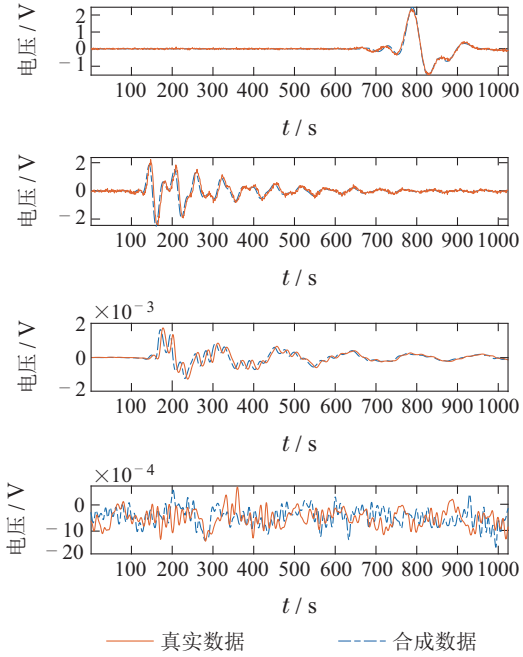
(f) TTS-GAN







(g) MAD-GANs



(h) TCGN

图 2　真实样本和生成样本的比较

Fig. 2　Comparison of real and generated samples

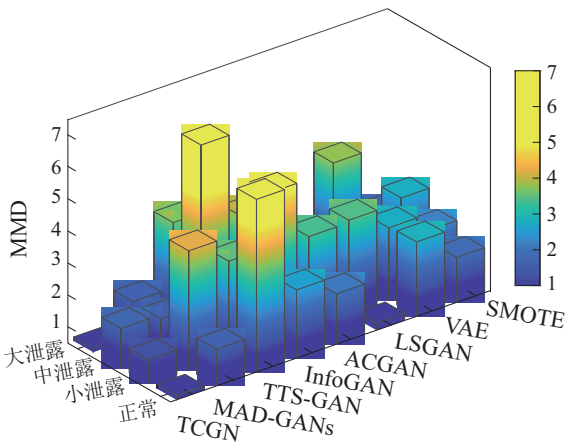


图 3　真实分布和生成分布之间的统计学相似性

Fig. 3　Statistical similarity between real and synthetic distributions

在实验中, 正常状态的管道数据量为2 000, 其他健康状态的数据量随AR变化. 数据集的详细信息如表3所示.

表 3 数据集信息
Table 3 Information of dataset

AR	50:1	20:1	10:1	1:1
大泄漏	40	100	200	2 000
中泄漏	40	100	200	2 000
小泄漏	40	100	200	2 000
正常状态	2 000	2 000	2 000	2 000

为了减小分类模型对诊断性能的影响, 本文选择了适用于时间序列分析的LSTM作为故障分类器. 训练过程中, 批量大小为256, 训练次数为5 000, 优化器为Adam, 学习率为1e–4.

4.2 评价指标

根据文献 [21–22], 本文采用精确率、召回率、F1分数、特异性作为故障诊断性能评估指标.

4.3 故障诊断实验结果

表4和表5展示了所有方法的实验结果. 表中所有最高准确率以粗体标注, 所有次优准确率以下划线标注. 表4给出了$\mathrm{AR}=50:1$和$\mathrm{AR}=20:1$情况下的实验结果. 在$\mathrm{AR}=50:1$情况下, TCGN算法优于所有比较的传统算法, 并在深度方法中获得了第2好的诊断结果, 有最佳的召回率(72.50%)以及第2好的精确率 (74.61%)、F1分数 (73.49%) 和特异性 (92.26%), 这接近于 TTS-GAN 所取得的最好的精确率, F1分数(+2.3%)和特异性.

表 4 增强比例为50:1和20:1的不同合成数据集的故障诊断性能(%)(括号内为提升比率)
Table 4 Diagnostic performance of different synthetic datasets for AR 50:1 and 20:1 (%)

	50:1				20:1			
	精确率	召回率	F1分数	特异性	精确率	召回率	F1分数	特异性
SMOTE	74.16(+0.6)	69.05(+4.9)	71.38(+2.9)	89.37(+3.2)	73.53(+33.5)	72.06(+6.7)	72.68(+9.8)	92.02(+1.8)
VAE	74.08(+0.7)	69.10(+4.9)	71.39(+2.9)	86.03(+7.2)	73.38(+33.8)	72.22(+6.5)	72.69(+9.8)	89.99(+4.1)
LSGAN	74.27(+0.5)	67.31(+7.7)	70.08(+4.8)	88.28(+4.5)	73.36(+33.8)	73.61(+4.5)	73.44(+8.7)	88.49(+5.8)
ACGAN	72.54(+2.8)	64.48(+12.4)	67.11(+9.5)	89.55(+3.0)	73.65(+33.8)	75.00(+2.6)	74.31(+7.4)	91.67(+2.2)
InfoGAN	74.46(+0.2)	68.45(+5.9)	71.01(+3.5)	88.71(+4.0)	71.27(+37.7)	72.72(+5.8)	71.87(+11.1)	89.29(+4.9)
TTS-GAN	**81.90**	70.83(+2.3)	**73.71**	**95.64**	84.01(+16.8)	75.16(+2.3)	74.75(+6.8)	92.83(+0.9)
MAD-GANs	74.31(+0.4)	67.81(+6.9)	70.46(+4.3)	89.76(+2.7)	72.89(+34.7)	75.00(+2.6)	73.92(+8.0)	91.73(+2.1)
TCGN	74.61	**72.50**	73.49	92.26	**98.19**	**76.95**	**79.85**	**93.70**

表 5 增强比例为10:1和1:1的不同合成数据集的故障诊断性能(%)(括号内为提升比率)
Table 5 Diagnostic performance of different synthetic datasets for AR 10:1 and 1:1 (%)

	10:1				1:1			
	精确率	召回率	F1分数	特异性	精确率	召回率	F1分数	特异性
SMOTE	72.99(+35.0)	75.00(+13.1)	73.95(+19.4)	92.52(+2.3)	81.05(+21.7)	80.96(+21.8)	80.84(+22.0)	93.84(+6.1)
VAE	72.69(+35.5)	73.47(+15.4)	73.00(+21.0)	92.14(+2.8)	74.90(+31.7)	74.91(+31.7)	74.75(+32.0)	91.70(+8.5)
LSGAN	91.29(+7.9)	74.89(+13.2)	77.10(+14.5)	91.74(+3.2)	84.61(+16.6)	84.54(+16.7)	84.45(+16.8)	94.87(+4.9)
ACGAN	92.63(+6.3)	76.36(+11.0)	76.86(+14.9)	92.14(+2.8)	88.19(++11.8)	88.18(+11.9)	88.18(+11.9)	98.26(+1.3)
InfoGAN	98.16(+0.3)	78.31(+8.2)	80.40(+9.8)	93.20(+1.6)	90.15(+9.4)	90.14(+9.4)	90.13(+9.4)	96.91(+2.7)
TTS-GAN	98.34(+0.2)	81.25(+4.3)	84.14(+4.9)	93.71(+1.0)	93.75(+5.2)	93.63(+5.4)	93.67(+5.3)	97.92(+1.6)
MAD-GANs	96.62(+1.9)	76.72(+10.5)	76.47(+15.5)	91.64(+3.3)	92.59(+6.5)	92.33(+6.8)	92.32(+6.8)	97.40(+2.2)
TCGN	**98.55**	**84.80**	**88.34**	**94.73**	**98.67**	**98.69**	**98.68**	**99.57**

需要注意的是, 作为一种基于深度神经网络的方法, TTS-GAN显示出相当有竞争力的性能. 此外, 在$\mathrm{AR}=20:1$情况下, TCGN算法不仅优于传统算法, 而且获得了比所有深度方法更好的诊断结果, 在精确率、召回率、F1分数, 以及特异性上均取得了最高的准确率.

表5显示了$\mathrm{AR}=10:1$和$\mathrm{AR}=1:1$情况下的实验结果. 在$\mathrm{AR}=10:1$情况下, 相较于表4中的实验结果, 在数据集趋于平衡的过程中, TCGN取得了很大的进步, 在精确率(98.55%)、召回率(84.80%)、F1分

数(88.34%)和特异性(94.73%)方面都获得了最好的性能, 超过了第2好的TTS-GAN算法, 分别提升了0.2%, 4.3%, 4.9%, 1.0%. 此外, 在完全均衡的数据集下, 基于TCGN算法的分类模型性能得到了显著的提升, 取得了最佳的诊断结果.

从表4和表5的实验结果中可以看出, TCGN算法的综合性能优于所有对比算法, 并在不同的AR下表现出了更强的鲁棒性. 结论如下: 1) 与基于过采样的SMOTE方法以及基于重构误差的VAE方法相比, 基于对抗框架的深度生成模型可以更好地解决小样本和类不平衡问题; 2) 未考虑时间结构和类别信息的传统GAN算法及其变体, 仅注重通过深层网络提取特征并合成样本. 然而, 这些特征的时间结构不清晰且故障属性较弱, 很难提升故障诊断准确率; 3) 更高的精确率代表着更低的误报率, 更高的召回率代表着更低的漏报率. 在AR = 50 : 1, 20 : 1, 10 : 1的情况下, TCGN算法一直呈现高精确率、低召回率的特征. 此时分类模型虽无法准确检测所有健康状态的所有数据, 但是分类模型得到的阳性预测结果可信度非常高. 随着数据集趋于平衡, TCGN算法表现出了高精确率、高召回率的特征, 此时分类效果最好.

5　讨论与性能分析

为了更好地评估本文的方法, 本文从特征可视化以及消融研究方面分析其性能.

5.1　特征可视化

为了进一步验证合成样本故障特征的有效性, 本文采用T-SNE算法将不同健康状态下的管道数据降维并投影至二维平面, 以衡量从真实样本和合成样本中提取的特征的相似性. 从图4(a)–(b)中可以看出, SMOTE算法与VAE算法混淆了不同健康状态的样本特征, 导致合成数据散落在全局区域内, 既不靠近同类别真实数据也不区别于不同类合成数据. 另外, 由图4(c)–(e)可知, 传统GAN算法及其现有变体也不能很好地保留样本的判别特征. 其次, 通过观察图4(f)–(g)可以发现, 尽管合成数据有同类聚拢和异类分散的趋势, 但对应健康状态下的合成数据与真实数据之间仍存在误匹配问题. 对比上述实验结果, 本文的TCGN算法展现了更为优越的性能, 如图4(h)所示.

5.2　消融研究

为了确认所提方法中每个组成部分对性能提升的贡献, 本文在本小节进行了消融研究. TCGN算法的变体定义如下: 1) TCGN-TTC: 从TCGN模型中删除了TTC损失; 2) TCGN-CAC: 从TCGN模型中删除了CAC损失; 3) TCGN-Markov: 采用了传统优化策略代替基于马尔可夫链的多模态切换策略来训练TCGN模型.

通过观察图5, 可以得出以下结论: 1) TTC损失可以学习真实样本的时间动态特性, 从而提高合成样本时间结构的真实度和减弱噪声干扰; 2) CAC损失可以聚拢同类别真实样本与合成样本, 并扩大非同类真实样本与合成样本之间的距离; 3) 基于马尔可夫链的多模态切换策略可以动态调整不同损失之间的平衡, 保证TCGN模型可以沿着最优路径完成所有学习任务, 并快速达到收敛状态; 4) 只有当所有组成部分同时工作时, TCGN算法才可以达到最佳生成性能.

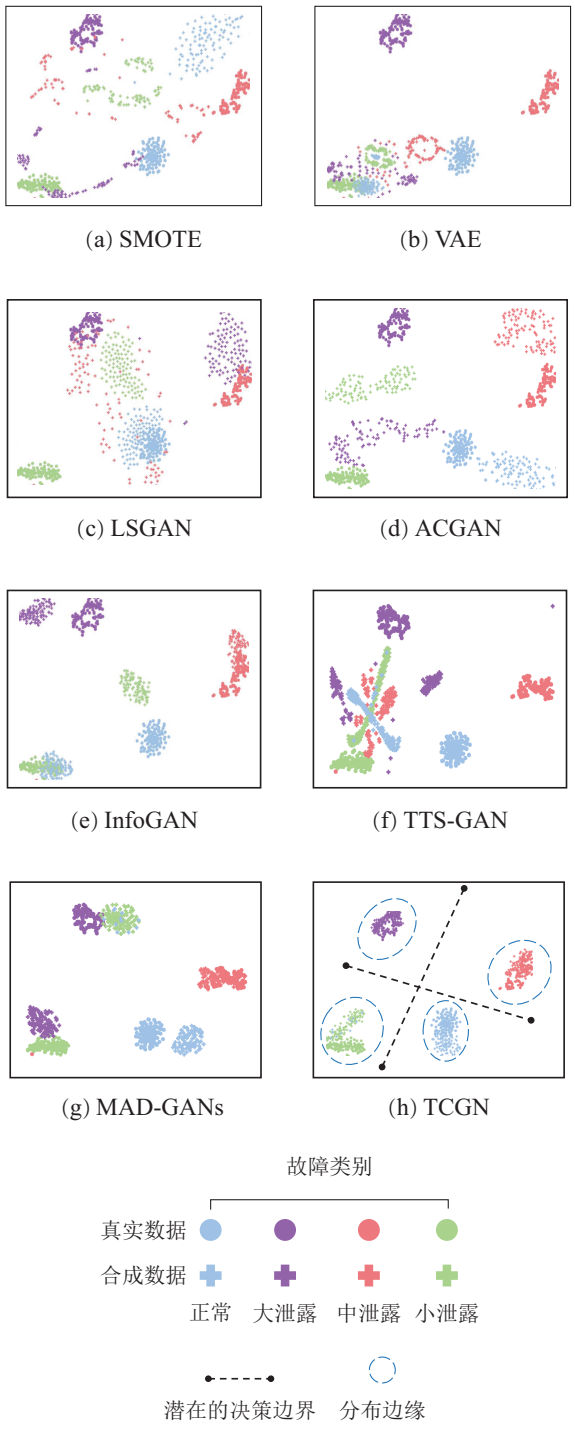
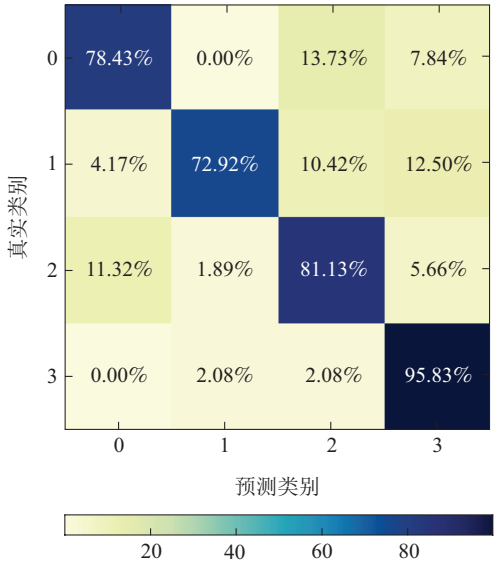
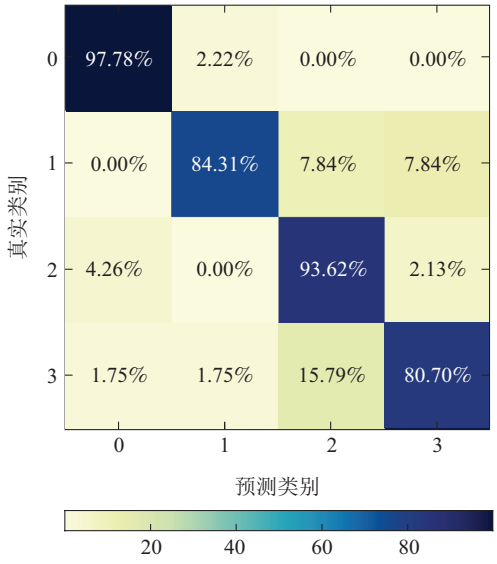


图 4　T-SNE特征可视化

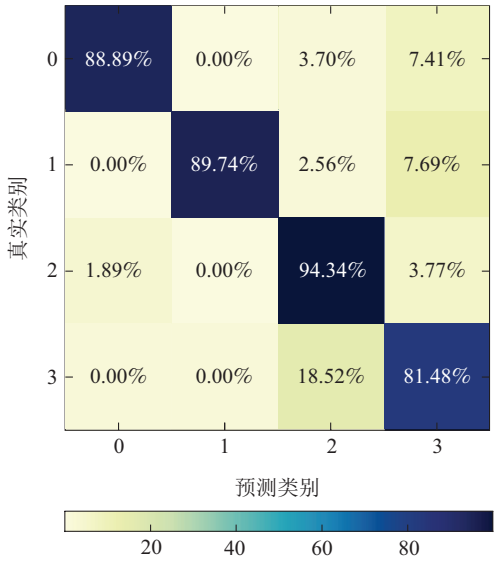
Fig. 4　Feature visualizations via T-SNE



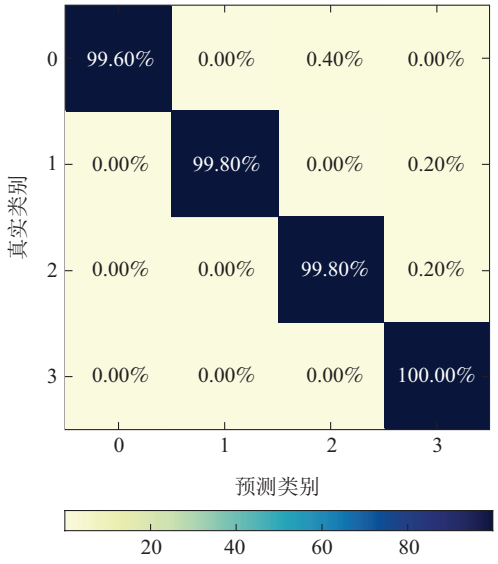
(a) TCGN-TTC (82.08%)



(b) TCGN-CAC (89.10%)



(c) TCGN-Markov (88.61%)



(d) TCGN (99.80%)

图 5 每种方法的混淆矩阵

Fig. 5 Confusion matrix for each method

6 结论

本文提出了一种基于马尔可夫链的多模态时序对比生成模型. 首先, 为了提高实例层面的真实性, 提出了TTC损失以增强合成数据与真实数据在时间维度的相似度. 此外, 考虑到合成分布与真实分布之间存在的误对齐问题, 设计了CAC损失, 以促进合成数据向同类别的真实数据聚拢, 分散不同类别的数据. 在此基础上, 引入了一种基于马尔可夫链的多模态切换策略, 实现了TCGN算法在生成、刻画、探索、收敛4个模态之间的自适应切换优化, 有效提高了合成数据的质量. 最后, 将所提出的TCGN算法应用于管道数据增强, 成功解决了小样本和类不平衡问题. 实验结果表明, TCGN算法在视觉评估和量化指标方面均优于一些已存在的数据增强算法, 显著提升了管道故障诊断准确率.

参考文献:

[1] WANG C, WANG Z, HAN Q, et al. Novel leader-follower-based particle swarm optimizer inspired by multiagent systems: Algorithm, experiments, and applications. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 2023, 53(3): 1322 – 1334.

[2] ZHANG H, HU X, MA D, et al. Insufficient data generative model for pipeline network leak detection using generative adversarial networks. *IEEE Transactions on Cybernetics*, 2020, 52(7): 7107 – 7120.

[3] LEI Y, YANG B, JIANG X, et al. Applications of machine learning to machine fault diagnosis: A review and roadmap. *Mechanical Systems and Signal Processing*, 2020, 138: 106587.

[4] JIN Jiangtao, XU Zifei, LI Chun, et al. Rolling bearing fault diagnosis based on deep learning and chaotic feature fusion. *Control Theory & Applications*, 2022, 39(1): 109 – 116.
(金江涛, 许子非, 李春, 等. 基于深度学习与混沌特征融合的滚动轴承故障诊断. 控制理论与应用, 2022, 39(1): 109 – 116.)

[5] GOODFELLOW I J, ABADIE J P, MIRZA M, et al. Generative adversarial nets. *Proceedings of Advances in Neural Information Processing Systems (NIPS)*. Red Hook, NY, USA: Curran, 2014: 2672 – 2680.

[6] ESTEBAN C, HYLAND S L, RÄTSCH G. Real-valued (medical) time series generation with recurrent conditional GANs. *Arxiv Preprint*, 2016, arXiv: 1706.02633.

[7] WANG Y, DONG X, WANG L, et al. Optimizing small-sample disk fault detection based on LSTM-GAN model. *ACM Transactions on Architecture and Code Optimization*, 2022, 19(1): 1 – 24.

[8] DONG X, GU C, WANG Z. Research on shape-based time series similarity measure. *2006 International Conference on Machine Learning and Cybernetics*. Piscataway: IEEE, 2006: 1253 – 1258.

[9] WANG C, WANG Z, LIU W, et al. A novel deep offline-to-online transfer learning framework for pipeline leakage detection with small samples. *IEEE Transactions on Instrumentation and Measurement*, 2023, 72: 3503913.

[10] MIRZA M, OSINDERO S. Conditional generative adversarial nets. *Arxiv Preprint*, 2014, arXiv: 1411.1784.

[11] ZENG N, WANG Z, LIU W, et al. A dynamic neighborhood-based switching particle swarm optimization algorithm. *IEEE Transactions on Cybernetics*, 2020, 52(9): 9290 – 9301.

[12] WANG C, XU C, YAO X, et al. Evolutionary generative adversarial networks. *IEEE Transactions on Evolutionary Computation*, 2019, 23(6): 921 – 934.

[13] CHAWLA N V, BOWYER K W, HALL L O, et al. SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 2002, 16(1): 321 – 357.

[14] KINGMA D P, WELLING M. Auto-encoding variational bayes. *Arxiv Preprint*, 2013, arXiv: 1312.6114.

[15] MAO X, LI Q, XIE H, et al. Least squares generative adversarial networks. *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. Piscataway, NJ: IEEE, 2017: 2794 – 2802.

[16] ODENA A, OLAH C, SHLENS J. Conditional image synthesis with auxiliary classifier GANs. *Proceedings of the International Conference on Machine Learning (ICML)*. Cambridge, MA: JMLR, 2017: 2642 – 2651.

[17] CHEN X, DUAN Y, HOUTHOOFT R, et al. InfoGAN: Interpretable representation learning by information maximizing generative adversarial nets. *Proceedings of Advances in Neural Information Processing Systems (NIPS)*. New York, USA: Curran Associates, Inc., 2016: 2180 – 2188.

[18] LI X, METSIS V, WANG H, et al. TTS-GAN: A transformer-based time-series generative adversarial network. *Arxiv Preprint*, 2022, arXiv: 2202.02691.

[19] LI D, CHEN D, JIN B, et al. MAD-GAN: Multivariate anomaly detection for time series data with generative adversarial networks. *Artificial Neural Networks and Machine Learning-ICANN 2019: Text and Time Series*. Cham: Springer, 2019: 703 – 716.

[20] WANG Chuang, HAN Fei, SHEN Yuxuan, et al. Full-Information particle swarm optimizer based on event-triggering strategy and its applications. *Acta Automatica Sinica*, 2023, 49(4): 891 – 903.
(王闯, 韩非, 申雨轩, 等. 基于事件触发的全信息粒子群优化器及其应用. 自动化学报, 2023, 49(4): 891 – 903.)

[21] WANG C, WANG Z, MA L, et al. Subdomain-alignment data augmentation for pipeline fault diagnosis: An adversarial self-attention network. *IEEE Transactions on Industrial Informatics*, 2023, DOI: 10.1109/TII.2023.3275701.

[22] CHEN X, ZHANG B, GAO D. Bearing fault diagnosis base on multi-scale CNN and LSTM model. *Journal of Intelligent Manufacturing*, 2021, 32: 971 – 987.

作者简介:

商　柔　博士研究生, 目前研究方向为深度学习与管道完整性分析, E-mail: shangrou61@126.com;

董宏丽　教授, 博士生导师, 目前研究方向为网络化控制系统、智能控制、传感器网络信息处理, E-mail: shiningdhl@vip.126.com;

王　闯　博士研究生, 目前研究方向为深度学习和迁移学习在管道故障诊断中的应用, E-mail: wangchuang64@126.com;

陈双庆　副教授, 目前研究方向为油气储运系统智能优化决策与控制, E-mail: csqing2590@163.com;

周国强　研究员, 目前研究方向为海洋平台结构监测与评价, E-mail: zhouguoqiang@263.net;

管　闯　讲师, 目前研究方向为智能故障诊断分析, E-mail: guanchuang126@126.com;

闫天红　讲师, 目前研究方向为海洋平台结构监测与评价, E-mail: yantianhong82@126.com.