基于生成对抗网络的分布式光伏窃电数据增强方法

李景歌,荣　娜,陈庆超

(贵州大学电气工程学院,贵州 贵阳 550025)

摘　要:由于分布式光伏窃电的稽查难度大,致使相关部门收集的窃电样本数量有限,无法满足基于数据驱动的窃电检测需求。通过数据增强的方式,提出一种基于 Wasserstein 生成对抗网络(WGAN)的分布式光伏窃电样本数据增强方法。首先,WGAN 通过生成网络与判别网络的对抗训练,能够学习到光伏窃电数据序列难以显式建模的时间相关性,可以生成与真实窃电样本具有相近分布的新的窃电样本;然后,根据典型的光伏窃电模型,针对窃电样本的数据特征选用卷积神经网络(CNN)进行窃电检测;最后,通过算例分析,对比不同数据增强方法与分类器,表明 WGAN 生成的窃电样本能够符合真实样本的波动规律和历史数据的概率分布特征,进而有效改善分类器的检测性能。

关　键　词:深度学习;生成对抗网络;数据增强;分布式光伏发电;窃电

DOI:10.19781/j.issn.1673-9140.2022.05.020　　**中图分类号**:TM615　　**文章编号**:1673-9140(2022)05-0181-10

A data augmentation method for distributed photovoltaic electricity theft using generative adversarial network

LI Jingge, RONG Na, CHEN Qingchao

(College of Electrical Engineering, Guizhou University, Guiyang 550025, China)

Abstract: Due to the difficulty of the inspection of distributed photovoltaic (PV) electricity theft, the number of electricity theft samples collected by relevant departments is limited, which cannot meet the needs of data-driven electricity theft detection. This paper proposes a data augmentation method for distributed PV electricity theft using Wasserstein generative adversarial network (WGAN). First, WGAN can explicitly learn the time correlation that is difficult to model in the PV electricity theft data sequence. Furthermore, it can generate new electricity theft samples with similar distributions to the real ones through the confrontation training of the generator and discriminator networks. Then, according to the typical PV electricity theft model and data characteristics, the convolutional neural network (CNN) is selected for electricity theft detection. Finally, through the case analysis, it is shown that the electricity theft samples generated by WGAN can conform to the fluctuation law of authentic samples and the probability distribution characteristics of historical data, thereby effectively improving the detection performance.

收稿日期:2021-08-14;**修回日期**:2021-10-07

基金项目:贵州省科学技术基金(2021277)

通信作者:荣　娜(1979—),女,博士,讲师,主要从事电力电子装备与系统、电力市场研究;E-mail:582969760@qq.com

Key words: deep learning; generative adversarial network; data augmentation; distributed photovoltaic power generation; electricity theft

近年来,新能源发电已成为各国在能源领域研究的重点。为推动分布式光伏发电项目发展,中国对光伏上网电量进行政策补贴[1-2]。由于补贴取决于用户发电量,导致部分用户通过某些技术手段来使得光伏上网电表多计用户发电量,进而获取更多补贴,该行为被称为光伏窃电行为。该行为给国家造成了巨大的经济损失,同时,用户为窃电改接线路存在着安全隐患。因此,相应的窃电检测研究对于新能源行业发展意义重大[3]。

目前,有监督分类和无监督回归 2 种主要窃电检测方法[4]。其中,无监督回归方法是基于光伏发电实测值与预测值的偏差值来识别和检测窃电行为,但该方法的检测精度低[5-6]。有监督分类方法主要包括传统的机器学习和深度神经网络,传统的机器学习包括:支持向量机(support vector machine,SVM)、极限梯度提升算法(extreme gradient boosting,XGBoost)等传统机器学习分类方法[7-8]。相对于深度神经网络,其算法相较简单,易于解释,数据量要求低,但其特征提取能力较低。目前深度神经网络已广泛应用于数据识别分类,常用模型有 BP 神经网络(back propagation neural network,BPNN)、卷积神经网络(convolutional neural network,CNN)等[9-11]。由于深度神经网络能够映射繁杂的非线性关系,并且兼备强大的特征提取能力,从而能够满足窃电样本的数据特征提取要求,比传统机器学习方法具有更高的检测准确率。

然而,在实际工程中,由于稽查难度大,从而收集的分布式光伏窃电样本过少,难以满足深度神经网络对样本数据量的要求。针对上述的问题,在数据层面,可以对样本量较少的数据集进行数据增强,进而将窃电检测精度提高。当前只有少数研究考虑到了数据增强问题,文献[12]采用合成少数类过采样技术(synthetic minority oversampling technique,SMOTE)通过数据交叉合成新样本,并且由于仅利用局部先验分布信息,未增加有效分类特征,从而易造成模型的过拟合;文献[13]利用变分自动编码器(variational auto-encoder,VAE)生成新样本,但在生成新样本的过程中,产生样本中未有的噪声,无法生成高质量样本,从而提升分类器的性能有限;文献[14]构建生成式对抗网络(generative adversarial network,GAN),通过对抗训练来学习样本数据的潜在分布,生成高质量的合成样本,解决了因客观因素导致的光伏发电功率等时间序列样本量不足问题。在文献[15]提出 GAN 后,GAN 逐渐已经在计算机视觉领域展现出了强大的数据生成能力,已成为学术界广为认可的一种生成模型[16]。GAN 在图像修复、图像生成、数据生成等领域的成功应用,证明了其通过无监督学习,能够学习到真实样本潜在分布特征[17-18]。因此,GAN 可以学习得到窃电样本的分布特征,生成高质量的窃电样本,但 GAN 在分布式光伏窃电样本数据生成的应用仍处于初级阶段。

综上所述,本文提出一种基于 Wasserstein 生成对抗网络(wasserstein generative adversarial network,WGAN)的分布式光伏窃电样本数据增强方法。设计由深度神经网络构成的生成器与判别器网络结构,通过 Wasserstein 距离约束生成窃电样本的损失,采用无监督学习充分挖掘窃电样本的潜在分布特征,可以生成服从窃电样本分布特征的新样本,来对窃电样本进行数据增强,提高窃电检测精度。

1 WGAN 的基本原理

1.1 GAN 的原理

生成对抗网络由生成器 G 与判别器 D 这 2 个深度神经网络网络构成。其中生成器的输入是服从某种分布的随机噪声微量 $\boldsymbol{z}$(一般服从正态分布,记作 p_z);生成器用于充分挖掘出窃电样本潜在的真实分布特征,并生成符合真实分布特征的新样本;将原始窃电样本 x 和生成的新样本 $G(\boldsymbol{z})$ 组成新的数据组送入判别器;判别器的任务是去判别出其输入属于真实样本还是生成器生成的新样本[19],其典型结构如图 1 所示。

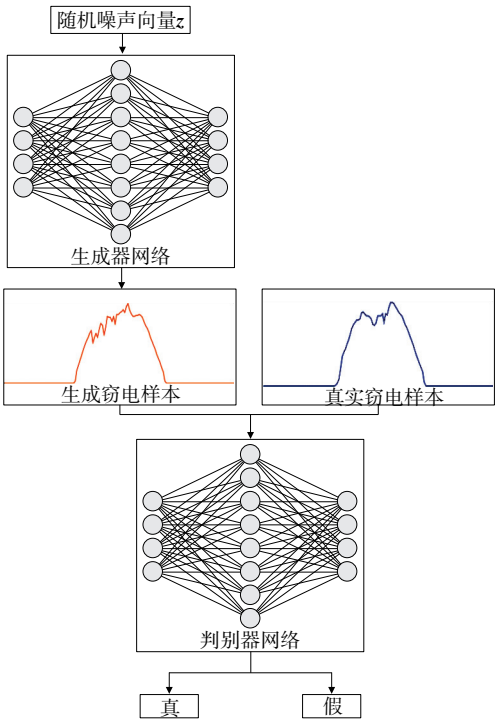


图 1　GAN 的典型结构

Figure 1　Typical structure of GAN

假设已收集了 n 组分布式光伏窃电样本，则窃电样本数据集可表示为 $X=\{x_1,x_2,\cdots,x_n\}$，其中 n 为 X 中窃电样本的总数目。窃电样本 x 服从某种潜在的分布特征 p_r，将服从高斯分布的随机噪声向量 $\boldsymbol{z}$ 输入到生成器中，然后输出满足 p_r 分布的新样本，通过构建深度神经网络，得到从 p_z 到 p_r 的映射关系。

生成器 G 为了让判别器 D 误以为其输入中 $G(\boldsymbol{z})$ 是真实的窃电样本，而不是生成器 G 生成的样本。生成器 G 就要通过无监督学习，尽量生成接近窃电样本分布的新样本。其损失函数被定义为 $-E_{z\sim p_z}[D(G(\boldsymbol{z}))]$，函数值越小表示新样本越服从 p_r 分布，生成新样本的质量越高。因此，生成器的任务是其损失函数最小化，即

$$\min\{-E_{z\sim p_z}[D(G(\boldsymbol{z}))]\} \tag{1}$$

判别器 D 的输入样本与数据集 X 中样本的维度一致，输入样本数据包括真实的窃电样本和生成器生成的窃电样本，判别器的任务是判别出其输入的是真实窃电样本还是生成器生成的新样本。判别器的输出是一个标量值，属于[0,1]，代表其输入的是真实样本或者生成样本的概率。D 的损失函数被定义为 $E_{x\sim p_r}[D(x)]-E_{z\sim p_z}[D(G(\boldsymbol{z}))]$，函数值越大，判别器鉴别出真伪的概率越大，所以训练目标是最大化 $E_{x\sim p_r[D(x)]}$ 和最小化 $E_{z\sim p_z}[D(G(\boldsymbol{z}))]$。因此，判别器的任务是最大化其损失函数，即

$$\max\{E_{x\sim p_r}[D(X)]-E_{z\sim p_z}[D(G(\boldsymbol{z}))]\} \tag{2}$$

综上所述，GAN 整体的训练过程是生成器和判别器之间的博弈问题，其中生成器生成尽量逼近真实窃电样本分布的新样本，减小生成样本与真实样本之间的分布差距，目的是骗过生成器。而判别器需要学习到样本间的分布差距以最大程度的将其辨别真伪[20]。其训练过程的目标函数为

$$\min_G\max_D\{E_{x\sim p_r}[D(x)]-E_{z\sim p_z}[D(G(\boldsymbol{z}))]\} \tag{3}$$

1.2　Wasserstein 距离改进的 GAN 原理

GAN 在训练过程中，当判别器接近最优时，由于 JS 散度会变成常数，致使梯度消失。同时，由于 GAN 的生成器和判别器在训练过程中，同步收敛困难，致使模式“崩溃”，导致其训练过程不稳定。因此 GAN 生成的样本不能满足高质量的需求。

WGAN 解决了原始 GAN 的训练过程的缺陷，减轻其训练过程中梯度消失的问题，提高训练稳定性[21]，从而易于生成高质量的样本。WGAN 用 Wasserstein 距离代替原始的 JS 散度，来衡量生成样本与真实样本之间分布的差距，但其原始 GAN 网络结构没有改变。Wasserstein 距离的定义为

$$W(p_r,p_g)=\inf_{\delta\in\prod(p_r,p_g)}E_{x,y\sim\delta}[\|x-y\|] \tag{4}$$

式中　$\prod(p_r,p_g)$ 为以 p_r 与 p_g 为边缘分布的联合概率分布 d 的集合；$W(p_r,p_g)$ 为生成分布 p_g 逼近真实分布 p_r 时，需要 x 向 y 移动的距离。

实际中，由于计算分布之间的 Wasserstein 距离非常困难，因此使用 $W(p_r,p_g)$ 的 Kantorovich-Rubinstein 对偶形式：

$$W(p_r,p_g)=K^{-1}\sup_{\|f\|_L\leqslant K}E_{x\sim p_r}[f(x)]-E_{x\sim p_g}[f(x)] \tag{5}$$

其中，$\|f\|_L\leqslant K$ 表明函数 $f(x)$ 满足 K-Lipschitz 连续条件，其导函数绝对值存在上界。

1.3　WGAN 的结构设计

WGAN 的生成器与判别器的训练方式是交替对抗训练，其中一次完整训练：先训练判别器 5 次，

再训练 1 次生成器。训练过程需要分开训练，当训练生成器时，固定判别器网络的权重，当训练判别器时，固定生成器的网络权重固定，使用 Adam 算法进行网络参数的更新。当达到一定的迭代次数时，损失函数接近于 0，所生成的样本的波动规律已经逼近真实窃电曲线，生成的新样本，可用于训练分类器。

结合分布式光伏窃电样本的数据特征，通过试验最终构建的网络结构：生成器网络使用 4 层全连接网络，其中前 3 层选用渗漏型线性整流函数(leaky rectified linear unit，LeakyReLU)为激活函数，最后 1 层选用双曲正切函数 tanh 为激活函数。判别器网络使用 3 层全连接网络，其中，前 2 层选用 LeakyReLU 为激活函数，最后 1 层选用二分类问题常用的非线性激活函数 sigmoid 为激活函数。

2 基于 WGAN 的分布式光伏窃电检测

2.1 分布式光伏窃电模型

分布式光伏电站的光伏阵列输出的直流电需要经过逆变器成为交流电，输出的交流电优先本地负荷使用，多余的电量通过并网传输到外部，简化原理如图 2 所示。其中，关口电表记录用户从外网获取的电量，而光伏计量表记录用户发电量。与传统窃电不同，分布式光伏窃电是为了骗取高额补贴。目前，存在的典型的光伏窃电方式如下[22]。

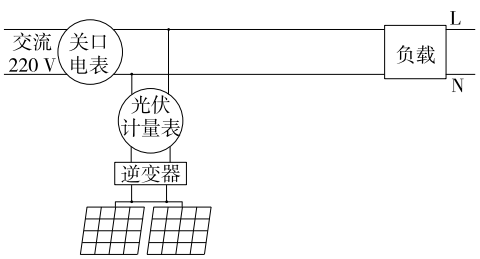


图 2 分布式光伏发电简化原理

Figure 2 Simplified principle of distributed PV system

1)升流法(升压法)。

当用户窃电时，利用电流发生器将工频电流加到光伏计量表的电流回路上，使电流高于正常值；当用户窃电时，利用升压变压器升高电压接入到光伏计量表的电压回路上，使电压高于正常值。用户使用这 2 种方法窃电，导致光伏计量表的测量功率高于正常值，从而多计发电量来骗取国家补贴。由于 2 种方法效果基本相同，本文归为一类窃电方式，这种窃电方式的数学模型如下：

$$\begin{cases} P'=(I+\Delta I)U=P(1+\dfrac{\Delta I}{I}) \\ P'=(U+\Delta U)I=P(1+\dfrac{\Delta U}{U}) \end{cases} \tag{6}$$

式中 P'为实际测量功率；P 为原本功率；I 为真实的电流；U 为真实的电压值；ΔI 为增加的电流；ΔU 为增加的电压值。

2)市电整流逆变法。

用户窃电时，利用电力电子装置冒充光伏阵列发电，将市电整流成直流电，再经过逆变器上网。使用此方式窃电，可以不安装光伏阵列，来骗取国家补贴。这种窃电方式的数学模型如下：

$$P'=P+\Delta P \tag{7}$$

式中 ΔP 为窃取的市电功率。

3)市电改接法。

窃电用户把市电的进线接到光伏计量表的进线，导致实际光伏计量表计量的是用户负载的用电量。使用该方法窃电时，窃电用户的功率曲线与正常用户的光伏计量表量测功率曲线差异较大。这种窃电方式的数学模型如下：

$$P'=\sum_{i=1}^{n}P_i \tag{8}$$

式中 P_i为第 i 种用户负载的用电量；n 为窃电用户负载的数量。

2.2 基于 CNN 的光伏窃电检测

卷积神经网络 CNN 具有强大的特征提取能力，已被成功应用于数据识别分类、故障诊断、语音识别等领域[23-24]。相较于与传统的机器学习分类，CNN 还具有强大的映射能力，能映射复杂的非线性关系，进而选用 CNN 为窃电检测的分类器。分类器 CNN 主要由卷积层、池化层、舍弃层、平坦层和全连接层组成。其卷积层和池化层用来提取数据特征；舍弃层随机以一定概率舍弃部分神经元，来缓解训练中的过拟合；平坦层起到格式转换的作用，将输入的矩阵数据转换成向量数据，即可连接最后的全连接层；全连接层来完成样本数据的分类。根据分布式光伏窃电样本的数据特征，通过试

验得到 CNN 的结构和参数，本文所使用网络结构与参数如表 1 所示。

表 1　卷积神经网络结构与参数

Table 1　Structure and parameters of CNN

网络层名称	参数	数值
卷积层 1	卷积核结构	2×2
	步长	1
	滤波器数目	16
	激活函数	ReLU
池化层	池化窗口	2×2
	步长	1
舍弃层 1	舍弃比率	0.25
卷积层 2	卷积核结构	2×2
	步长	1
	滤波器数目	32
	激活函数	ReLU
舍弃层 2	舍弃比率	0.25
全连接层 1	神经元数目	10
	激活函数	ReLU
全连接层 2	神经元数目	4
	激活函数	Softmax

以 1 h 为采样时间分辨率，一天 24 个样本点为例，阐述 CNN 中数据流的变化过程，根据 CNN 数据处理要求，将样本向量的第 24 维数据复制并拓展到第 25 维数据，并将 25 维的样本向量通过 Python 中的 Reshape 函数变为 5×5×1 的 3 维张量作为卷积层的输入向量。输入向量经过第 1 个卷积层连接池化层，随后经过第 1 个舍弃层再输入到第 2 个卷积层，再通过第 2 个舍弃层，最后经过平坦层连接 2 层神经元数量分别是 10 和 4 的全连接层，完成四分类任务，即正常、升流法（升压法）窃电、市电整流逆变法窃电和市电改接法窃电。其中 2 个卷积层和 1 个池化层提取数据特征，其卷积核和池化窗口的尺寸大小均为 2×2，舍弃层的舍弃率为 0.25，仅最后一层激活函数选择归一化指数函数（soft maximum，Softmax），其余的激活函数选用线性整流函数（rectified linear unit，ReLU）。网络中，选用多分类交叉熵函数为损失函数，选用可以对学习率进行自适应约束的 Adadelta 算法为优化器。

2.3　基于 WGAN 的窃电检测流程

2.3.1　模型的数据处理

光照强度等天气因素与分布式光伏发电的输出功率有很强相关性，由于早晚的光照强度相对较弱进而输出功率较小，导致一天中最大功率值与最小功率值的差值很大。由于市电改接法的光伏计量表测量的是用户耗电功率，所以测量功率受天气等因素较小。相对而言，另外 2 种窃电方法的光伏计量表测量功率受天气等因素较大。为消除天气等因素对分类器精度的影响，使用采样点光伏预测值和测量值的偏差作为其输入，数学表达式如下：

$$X=[\Delta P_1,\Delta P_2,\cdots,\Delta P_n] \tag{9}$$

式中　X 为分类器的输入变量；n 为样本采样的次数；ΔP_n 为第 n 次采样时的偏差值。

为保证模型训练的稳定性，需要对样本数据进行归一化减少数据差异，采用的归一化方法为

$$x_i'=\frac{x_i-x_{i,\min}}{x_{i,\max}-x_{i,\min}} \tag{10}$$

式中　x_i'、x_i 分别为数据处理前后的第 i 种变量；$x_{i,\max}$、$x_{i,\min}$ 分别为第 i 种变量的最大值和最小值；i 的值属于 $[1,n]$；n 为变量种类总数。

2.3.2　窃电检测过程具体步骤

1）偏差值计算。利用各类气象环境因素和光伏发电功率作为输入和输出，来训练循环神经网络预测模型[25]。随后将训练好的循环神经网络用来预测各时间采样点的功率，从而计算预测值和测量值的偏差用于窃电检测。

2）WGAN 的训练。训练开始时，由于生成器生成的窃电样本远离真实窃电样本的分布，判别器容易辨别真伪；随着训练次数增加进行，生成的新样本逐渐接近真实窃电样本分布；当生成器与判别器之间互相博弈达到纳什平衡时，损失函数逼近于 0，此时模型训练完成。

3）CNN 的训练。将 WGAN 产生的样本和真实窃电样本共同构建数据集。其中，CNN 的卷积层和池化层用来提取数据特征，通过误差反向传播算法更新网络权重。

4）模型性能评估。用准确率作为评价指标过于单一化。本文中四分类问题可以转化为 4 个二分类问题，除了准确率以外，还选用几何平均（geometric mean，G-mean）和宏平均 F_1 分数（macro-averaged F_1 score，Macro F_1）来评估 CNN 的性能[26]。其中，G-mean 和 Macro F1 指标数值越大，效果越好，其数学计算式如下：

$$P_i=\frac{n_{ii}}{\sum_{j=1}^{k}n_{ji}} \tag{11}$$

$$R_i=\frac{n_{ii}}{\sum_{j=1}^{k}n_{ji}} \tag{12}$$

$$F_1=\frac{\sum_{i=1}^{k}n_{ii}}{\sum_{i=1}^{k}\sum_{j=1}^{k}n_{ji}} \tag{13}$$

$$F_2=\left(\prod_{i=1}^{k}R_i\right)^{\frac{1}{k}} \tag{14}$$

$$F_3=\frac{1}{k}\sum_{i=1}^{k}\frac{2P_iR_i}{P_i+R_i} \tag{15}$$

式(11)～(15)中 F_1、F_2、F_3 分别为准确率、G-mean和 Macro F_1 指标；P_i、R_i 分别为第 i 个种类的精确率和召回率；n_{ij} 为把种类 i 误诊断成种类 j 的样本个数。

3 算例分析

3.1 实验硬件和数据

本实验的运行环境是开源的 VS Code 平台，选用深度学习框架是 tensorflow 1.9.0 和 keras 2.2.0；实验硬件为 11th Gen Intel(R) Core(TM) i5-11300H @ 3.10 GHz 的处理器，16 GB 的内存容量。

为了验证模型性能，算例分析数据选自美国再生能源实验室的光伏数据集[27]。该数据集以 1 h 为时间分辨率，采集光伏发电量以及对应的各种气象数据，其时间包含 2016 年 1 月 1 日到 2018 年 12 月 31 日。为了评估各种因素和光伏发电量之间的相关性强弱，利用皮尔逊相关系数选出强相关因素作为输入变量，用于训练循环神经网络。

通过本文典型光伏窃电模型得到不同类型的窃电样本。各种类型的样本数量如表 2 所示。

表 2 各种样本类型的样本数量

Table 2 Number of samples for various sample types

样本类型	总样本/个	训练样本/个	测试样本/个
正常	273	204	69
升流法(升压法)	273	204	69
市电整流逆变法	273	204	69
市电改接法	273	204	69

3.2 WGAN 的性能表现

WGAN 的训练过程的稳定性可以通过其训练过程中损失函数的数值变化来看，可视化损失函数如图 3 所示。由图 3 可知，当训练次数达到 300 次时，损失函数的数值已基本稳定不变，这表明训练后的 WGAN 已经收敛。

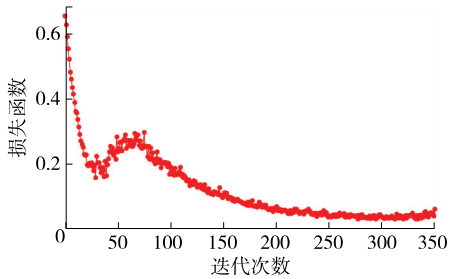


图 3 损失函数

Figure 3 Loss function

训练完 WGAN 后，用固定权重的生成器，输入高斯噪声生成不同类型样本，得到 816 个新的窃电样本，加入初始训练集使其样本数量翻倍。为验证其生成性能表现，如图 4 所示，对比测试样本中选取的部分样本与 WGAN 生成的窃电样本，来分析 WGAN 生成的窃电样本质量。

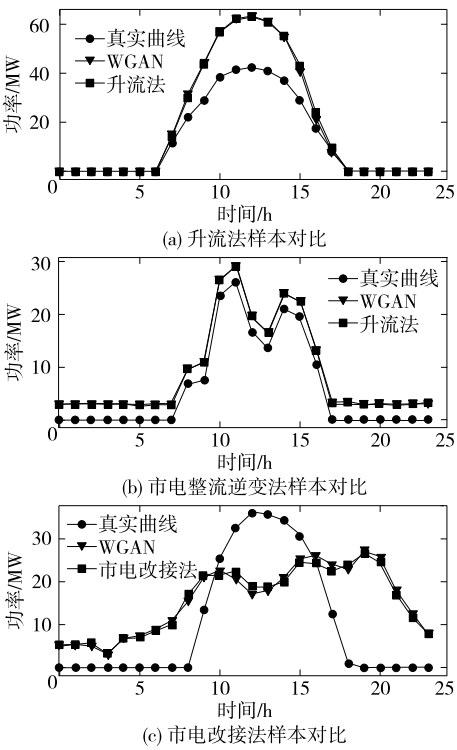
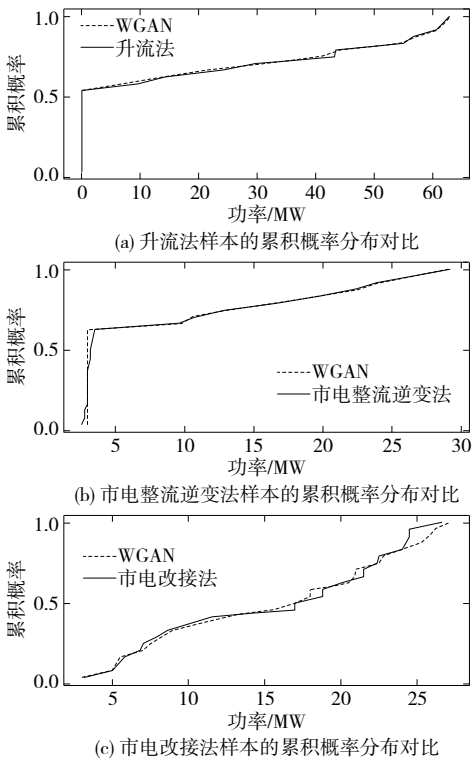


图 4 不同生成样本对比

Figure 4 Comparison of different generated samples

由图 4 可知，WGAN 产生的不同类型窃电样本的波动规律十分逼近真实的窃电样本。由于 WGAN 的训练过程中，测试集中的真实窃电样本并没有参与模型训练，表明 WGAN 能够通过无监督学习到符合真实样本的形状波动规律，较为符合实际情况。

仅从抽取样本的形状波动分析 WGAN 性能过于单一，进而选用累积分布函数，从数理统计角度来分析其生成性能表现，如图 5 所示。由图 5 可知，WGAN 生成的样本累积概率分布近似于真实的窃电样本，因此证明 WGAN 不仅兼顾窃电样本的形状，还能够学习到窃电样本潜在的概率分布特征，生成的高质量窃电样本符合真实窃电样本的分布特征。



(a) 升流法样本的累积概率分布对比

(b) 市电整流逆变法样本的累积概率分布对比

(c) 市电改接法样本的累积概率分布对比

图 5　不同生成样本的累积概率分布对比

Figure 5　Comparison of cumulative probability distributions of different generated samples

3.3　不同数据增强算法的有效性对比

为了验证数据增强生成窃电样本的有效性，以 CNN 为分类器，选择 SMOTE、GAN、VAE 和 WGAN 等方法进行对比实验。利用不同数据增强方法将初始训练集的样本数量增至原来 2 倍，然后对比 CNN 在数据增强前、后的性能表现，各项评估指标值如图 6 所示。

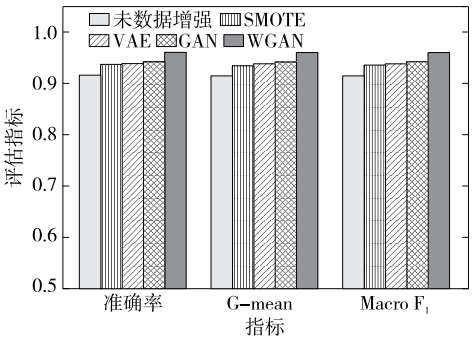


图 6　不同数据增强算法的对比

Figure 1　Comparison of different data augmentation methods

1）通过 SMOTE、VAE、GAN、WGAN 分别进行数据增强后，相较于未数据增强的训练集，CNN 的综合性能有不同程度的改善。对比使用不同算法进行数据增强后 CNN 分类器的性能表现，其中，使用 SMOTE 算法时，准确率、G-mean 和 Macro F_1 数值指标比未数据增强时分别提升了 2.11%、1.91%和 2.09%；使用 VAE 算法时，准确率、Macro F_1 和 G-mean 数值指标比未数据增强时分别提升了 2.23%、2.30%和 2.25%；使用 GAN 算法时，准确率、G-mean 和 Macro F_1 数值指标比未数据增强时分别提升了 2.63%、2.66% 和 2.63%；使用 WGAN 时，准确率、G-mean 和 Macro F_1 数值指标比未数据增强时分别提升了 4.43%、4.47% 和 4.42%。

2）相较于其他数据增强方法，SMOTE 对于分类器的综合性能的改善较差。这是由于该方法通过数据交叉合成新样本，仅利用局部先验分布信息，并未增加有效分类特征，从而易造成过拟合。

3）对比 3 种无监督学习，VAE 对于 CNN 性能的提升较差，这是因为 VAE 只能近似推理窃电样本的对数似然的下界，增加样本数量的同时也带来了原样本不具有的噪声。GAN 的训练过程会出现梯度消失和训练过程不稳定的问题，生成样本质量较低。而 WGAN 解决了 GAN 训练过程的缺陷，能够学习复杂数据的潜在分布，生成高质量的合成样本，对 CNN 性能的提升效果最明显。

3.4　WGAN 对不同分类器的性能提升对比

选择 SVM、XGBoost、BPNN 和 CNN 为分类器来进行对比实验，比较在 WGAN 进行数据增强前、

后分类器的表现，来验证 WGAN 生成样本的普适性。通过采用试探法得到不同分类器的结构和参数[28]，分类器 SVM 的内核选用径向基函数进行多分类模型训练；分类器 XGBoost 选用 gbtree 为分类的基学习器，设置其最大深度参数、采样比例参数、gamma 参数分别为 6、0.7、0.1；BPNN 的输入层和输出层的神经元个数分别是 24 和 4，中间隐含层由 2 层全连接层相连。对比在数据增强前、后不同分类器的性能表现，各项评估指标值如图 7 所示。

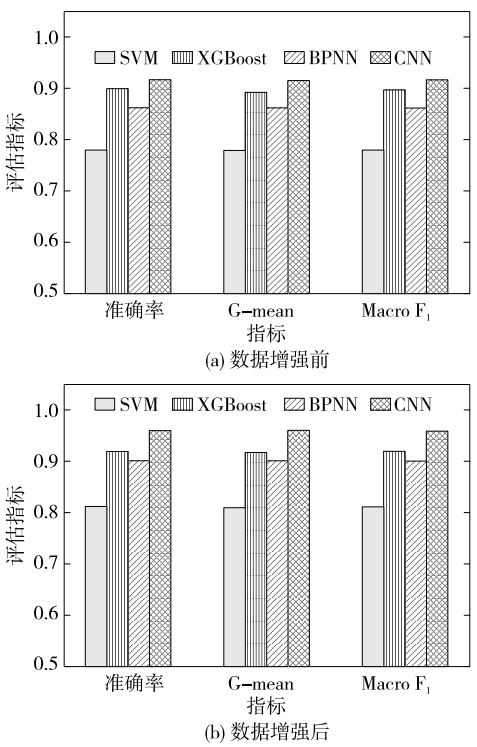


图 7 不同分类器的检测结果对比

Figure 7 Comparison of detection results of different classifiers

1）不同分类器的准确率、G-mean 和 Macro F_1 指标都很接近，这表明它们正确判别不同窃电样本类型的概率很接近。

2）通过 WGAN 进行数据增强后，各种分类器的综合性能得到显著改善。对比发现，SVM 分类器的准确率、G-mean 和 Macro F_1 数值指标比增强前分别提升了 3.30%、3.33%和 3.29%；XGBoost 分类器的准确率、G-mean 和 Macro F_1 数值指标比增强前分别提升了 2.14%、2.80%和 2.32%；BPNN 分类器的准确率、G-mean 和 Macro F_1 数值指标比增强前分别提升了 3.99%、4.09%和 4.05%；CNN 分类器的准确率、G-mean 和 Macro F_1 数值指标比增强前的训练集分别提升了 4.43%、4.47%和 4.42%。这说明 WGAN 能够学习生成高质量的窃电样本，显著提高基于数据驱动的分布式光伏窃电检测性能，能够适应不同的分类器。

4 结语

由于分布式光伏窃电隐蔽性强，电力公司掌握的分布式光伏窃电样本数量有限，本文提出了一种基于 WGAN 的分布式光伏窃电数据增强方法。WGAN 可以学习窃电样本难以显式建模的分布关系来生成符合真实样本分布的新样本；然后，根据典型光伏窃电模型，针对窃电样本的分布特征构建卷积神经网络来进行窃电检测；最后，通过算例分析，得出结论如下：

1）WGAN 解决 GAN 训练过程缺陷，通过对抗训练能够学习到真实窃电样本的分布特征，生成兼顾样本的形状和分布特征的新样本，该方法有效解决了数据量有限的问题；

2）将不同数据增强算法进行对比，使用 WGAN 数据增强后，CNN 分类器的准确率、G-mean 和 Macro F_1 指标表现最好，说明 WGAN 生成的样本质量最高，显著提升检测性能；

3）对比通过 WGAN 数据增强前后，不同分类器的准确率、G-mean 和 Macro F_1 指标都显著提升，说明 WGAN 的应用可以适应不同分类器，具有很强的适应性。

参考文献：

[1] 罗一凡，蒋传文，李春哲，等.考虑分布式电源入网的反窃电综合管理方法[J].电器与能效管理技术，2015(4)：49-55＋59.
LUO Yifan，JIANG Chuanwen，LI Chunzhe，et al. Integrated management method of anti-electricity stealing considering distributed generation[J]. Electrical & Energy Management Technology，2015(4)：49-55＋59.

[2] 薛阳，杨艺宁，廖文龙，等.基于非线性独立成分估计的分布式光伏窃电数据增强方法[J].电力系统自动化，2022，46(2)：171-179.
XUE Yang，YANG Yining，LIAO Wenlong，et al. Data

augmentation method for distributed photovoltaic electricity theft based on non-linear independent components estimation[J]. Automation of Electric Power Systems, 2022, 46(2): 171-179.

[3] 路艳巧，孙翠英，曹红卫，等.基于边缘计算与深度学习的输电设备异物检测方法[J].中国电力，2020，53(6)：27-33.
LU Yanqiao, SUN Cuiying, CAO Hongwei, et al. Foreign body detection method for transmission equipment based on edge computing and deep learning[J]. Electric Power, 2020, 53(6): 27-33.

[4] 贾亦敏，史丽萍，严鑫.基于精英混沌蜂群算法优化小波神经网络的变压器故障诊断[J].高压电器，2020，56(8)：230-236.
JIA Yimin, SHI Liping, YAN Xin. Transformer fault diagnosis using wavelet neural network based on elite-chaos artificial bee colony algorithm[J]. High Voltage Apparatus, 2020, 56(8): 230-236.

[5] LIU X, NIELSEN P S. Scalable prediction-based online anomaly detection for smart meter data [J]. Information Systems, 2018, 77: 34-47.

[6] JANETZKO H, STOFFEL F, MITTELSTÄDT S, et al. Anomaly detection for visual analytics of power consumption data [J]. Computers & Graphics, 2014, 38: 27-37.

[7] 巢政，温蜜.一种基于 SMOTE 和 XGBoost 的窃电检测方案[J].智慧电力，2020，48(11)：97-102.
CHAO Zheng, WEN Mi. Scheme for electricity theft detection based on SMOTE and XGBoost[J]. Smart Power, 2020, 48(11): 97-102.

[8] CHEN T, GUESTRIN C. Xgboost: a scalable tree boosting system[C]//The 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, California, USA, 2016.

[9] BUZAU MM, TEJEDOR-AGUILERA J, CRUZ-ROMERO P, et al. Hybrid deep neural networks for detection of non-technical losses in electricity smart meters [J]. IEEE Transactions on Power Systems, 2019, 35(2): 1254-1263.

[10] 董立红，肖纯朗，叶鸥，等.一种基于 CAEs-LSTM 融合模型的窃电检测方法[J].电力系统保护与控制，2022，50(21)：118-127.
DONG Lihong, XIAO Chunlang, YE Ou, et al. Electricity theft detection method based on a CAEs-LSTM fusion model[J]. Power System Protection and Control, 2022, 50(21): 118-127.

[11] 高嵩，陆倚鹏，王笑倩，等.基于深度学习的悬式瓷绝缘子红外图像识别方法[J].电力科学与技术学报，2020，35(5)：119-125.
GAO Song, LU Yipeng, WANG Xiaoqian, et al. Infrared image recognition method of porcelain disc-suspended insulators based on deep learning technology[J]. Journal of Electric Power Science and Technology, 2020, 35(5): 119-125.

[12] FIGUEROA G, CHEN Y S, AVILA N, et al. Improved practices in machine learning algorithms for NTL detection with imbalanced data[C]//IEEE Power & Energy Society General Meeting, Chicago, USA: IEEE, 2017.

[13] 黄南天，王文婷，蔡国伟，等.计及复杂气象耦合特性的模块化去噪变分自编码器多源-荷联合场景生成[J].中国电机工程学报，2019，39(10)：2924-2934.
HUANG Nantian, WANG Wentin, CAI Guowei, et al. The joint scenario generation of multi source-load by modular denoising variational autoencoder considering the complex coupling characteristics of meteorology [J]. Proceedings of the CSEE, 2019, 39(10): 2924-2934.

[14] 李远松，高博，须琳，等.基于差分序列方差与 CPS 融合的数字变电站数据异常检测方法[J].电网与清洁能源，2021，37(2)：30-41.
LI Yuansong, GAO Bo, XU Lin, et al. An anomaly detection method for digital substation abnormal data based on fusion of difference sequence variance and CPS[J]. Power System and Clean Energy, 2021, 37(2): 30-41.

[15] GOODFELLOW I, POUGET-ABADIE J, MIRZA M, et al. Generative adversarial networks[J]. Communications of the ACM, 2020, 63(11): 139-144.

[16] 陈佛计，朱枫，吴清潇，等.生成对抗网络及其在图像生成中的应用研究综述[J].计算机学报，2021，44(2)：347-369.
CHEN Foji, ZHU Feng, WU Qingxiao, et al. A survey about image generation with generative adversarial nets [J]. Chinese Journal of Computers, 2021, 44(02): 347-369.

[17] 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 Jianwei, XIE Haojie, LUO Xionglin. Research progress on application of generative adversarial networks in various fields[J]. Acta Automatica Sinica, 2020, 46(12): 2500-2536.

[18] GE L, LIAO W, WANG S, et al. Modeling daily load profiles of distribution network for scenario generation using flow-based generative network [J]. IEEE Access, 2020, 8: 77587-77597.

[19] 王守相,陈海文,潘志新,等.采用改进生成式对抗网络的电力系统量测缺失数据重建方法[J].中国电机工程学报,2019,39(1):56-64.
WANG Shouxiang, CHEN Haiwen, PAN Zhixin, et al. A reconstruction method for missing data in power system measurement using an improved generative adversarial network[J]. Proceedings of the CSEE, 2019, 39(1): 56-64.

[20] 廖一帆,武志刚.基于迁移学习与 Wasserstein 生成对抗网络的静态电压稳定临界样本生成方法[J].电网技术,2021,45(9):3722-3728.
LIAO Yifan, WU Zhigang. The method to generate static voltage stability critical sample based on transfer learning and wasserstein generative adversarial network[J]. Power System Technology, 2021, 45(9): 3722-3728.

[21] WANG Q, ZHOU X, WANG C, et al. WGAN-based synthetic minority over-sampling technique: Improving semantic fine-grained classification for lung nodules in CT images [J]. IEEE Access, 2019, 7: 18450-18463.

[22] 王晓琦.基于光伏出力计算模型的窃电监管技术研究[D].南京:东南大学,2016.

[23] 招景明,唐捷,潘峰,等.基于 SDAE 和双模型联合训练的低压用户窃电检测方法[J].电测与仪表,2021,58(12):161-168.
ZHAO Jingming, TANG Jie, PAN Feng, et al. Detection method of electricity theft for low-voltage users based on SDAE and double-model joint training[J]. Electrical Measurement & Instrumentation, 2021, 58(12): 161-168.

[24] 蓝金辉,王迪,申小盼.卷积神经网络在视觉图像检测的研究进展[J].仪器仪表学报,2020,41(4):167-182.
LAN Jinhui, WANG Di, SHEN Xiaopan. Research progress on visual image detection based on convolutional neural network[J]. Chinese Journal of Scientific Instrument, 2020, 41(4): 167-182.

[25] 杨龙,吴红斌,丁明,等.新能源电网中考虑特征选择的 Bi-LSTM 网络短期负荷预测[J].电力系统自动化,2021,45(3):166-173.
YANG Long, WU Hongbin, DING Ming, et al. Short-term load forecasting in renewable energy grid based on bi-directional long short-term memory network considering feature selection[J]. Automation of Electric Power Systems, 2021, 45(3): 166-173.

[26] 杨德昌,廖文龙,任翔,等.基于胶囊网络的电力变压器故障诊断[J].高电压技术,2021,47(2):415-425.
YANG Dechang, LIAO Wenlong, REN Xiang, et al. Fault diagnosis of transformer based on capsule network[J]. High Voltage Engineering, 2021, 47(2): 415-425.

[27] National renewable energy laboratory. Solar integration national dataset toolkit[EB/OL]. https://www.nrel.gov/grid/sind-toolkit.html, 2020-12-17.

[28] 廖文龙,于赟,王煜森,等.基于图卷积网络的配电网无功优化[J].电网技术,2021,45(6):2150-2160.
LIAO Wenlong, YU Yun, WANG Yusen, et al. Reactive power optimization of distribution network based on graph convolutional network[J]. Power System Technology, 2021, 45(6): 2150-2160.