面向云计算应用的用电负荷数据差分隐私保护方法

于 群[1],沈志恒[2],孙飞飞[2],李知艺[1]
(1. 浙江大学 电气工程学院,浙江 杭州 310027;2. 国网浙江省电力公司经济技术研究院,浙江 杭州 310016)

摘要:随着云计算技术的发展,用户可以利用公共计算资源低成本、高效率地完成机器学习等大数据分析业务,但在提升计算效率和经济效益的同时,也面临隐私泄露风险。针对以机器学习即服务为代表的云计算中潜藏的用电负荷数据泄露问题,提出了一种差分隐私保护框架下基于时序生成对抗网络的用电负荷数据脱敏方法,通过使用满足差分隐私的脱敏合成数据替代原始敏感数据,从而有效阻止攻击者根据窃取的训练数据推断真实的敏感信息。引入瑞利差分隐私机制,在保留负荷数据统计学特征的前提下去除个体特征;在此基础上,采用循环神经网络作为生成对抗网络的生成器和判别器,捕获负荷数据的动态时间特性;同时,将自编码器与生成对抗网络相结合,进一步挖掘负荷数据的静态特征。通过理论推导证明了所提方法能够满足差分隐私要求,且可以对总隐私预算进行量化。数值实验结果表明,所提方法能保证隐私保护处理后用电负荷数据的隐私性和可用性。

关键词:用电负荷数据;云计算;差分隐私保护;生成对抗网络;自编码器;数据脱敏
中图分类号:TP 311.13;TM 714 **文献标志码**:A DOI:10.16081/j.epae.202205059

0 引言

在新型电力系统的建设过程中,海量物联网设备的广泛部署产生了大量有价值的数据信息,基于机器学习的大数据分析方法在电网规划运行中的应用越来越普遍,在能源预测、稳定控制、故障诊断、市场运营等方面得到广泛应用[1]。与此同时,云计算作为新一代信息技术,将数据中心软硬件资源整合为虚拟计算资源。随着"东数西算"工程的全面启动,计算资源成为一种可灵活调用的基础公共资源[2]。在此背景下,用户或配电网运营商可以从云平台购买弹性计算资源,利用云计算服务低成本、高效率地完成各项大数据分析业务[3]。

机器学习即服务 MLaaS(Machine Learning as a Service)是云计算的一种具体应用模式,能为数据持有者提供基于机器学习的数据处理、模型训练、预测服务和部署等自动化解决方案,吸引机器学习实践者在云平台部署应用程序,而无需建立自身的大规模基础设施和计算资源[4]。在典型的基于云平台的机器学习体系结构中,云平台提供机器学习模型及接口,终端用户将训练数据集上传至云平台,云平台将机器学习的运行结果返回给终端用户。

虽然MLaaS技术可以提升用户的效率和经济效益,但是会对用户数据隐私造成潜在的威胁。用户用电数据中暗含了大量的敏感用户信息,通过非侵入式电力负荷监测 NILM(Non-Intrusive Load Monitoring)[5-6]等数据挖掘技术,可以推断用户用电设备的运行状态,进而获得用户的行为信息,包括普通用户的生活习惯、人口数量、经济状况以及高敏感度用户(如军工企业或重要科研单位)的装备产量、生产工艺、科研进展等。当用户将用电数据作为训练数据集上传云端时,一旦云平台被黑客攻破,将造成用户敏感信息的泄露。在人工智能时代,个人隐私保护愈发受到国内外的重视和关注。《中华人民共和国网络安全法》、欧盟《通用数据保护条例》等国内外法律法规的施行,对企业处理用户数据的行为提出了明确的要求。如何保护MLaaS技术中的数据隐私,保证训练数据中的个人敏感信息不会被未授权人员直接或间接获取,成为制约云计算广泛应用的重要因素。

为了避免用户真实数据被攻击者获取,在用户上传训练数据集至云平台之前,可先对数据进行脱敏处理,抹去用户的敏感信息并确保数据的可用性。此类在上传数据之前即对数据进行保护的方法,能从源头上增强隐私保护的可靠性,进而大幅降低敏感信息的泄露风险。目前,基于差分隐私DP(Differential Privacy)[7]的本地化隐私保护方法在电力系统中应用较多,通常是对多用户的聚合用电数据和区域总体用电统计信息进行扰动[8]。然而,这些方法大多直接向原数据集中添加噪声,使数据的可用性不可避免地受到损失。为了实现数据隐私性和可用性之间的平衡,有学者提出了一种基于生成模型的隐私保护方法[9],用合成数据替代真实数据用于

收稿日期:2022-03-28;**修回日期**:2022-05-17
在线出版日期:2022-05-24
基金项目:国家自然科学基金资助项目(U2066601);国网浙江省电力公司科技项目(B311JY21000B)
Project supported by the National Natural Science Foundation of China(U2066601) and the Science and Technology Program of State Grid Zhejiang Electric Power Company (B311JY21000B)

机器学习。生成对抗网络GAN(Generative Adversarial Network)[10-11]能够从少量训练数据中学习到真实的数据分布,生成难辨真伪的高质量合成样本,在电力系统中被广泛用于海量新能源场景生成[12]、缺失数据修复[13]、光伏功率短期预测[14]等研究领域。然而,多项研究证实了生成对抗网络易被攻击,攻击者可以从其生成的合成数据中推理重建训练样本[15],从而使合成数据失去隐私保护的作用。因此,如何在保证合成数据可用性的前提下保护真实数据的敏感信息,是亟待解决的问题。

本文针对云平台MLaaS中可能产生的敏感数据泄露问题,在差分隐私保护框架下提出了一种新的基于时序生成对抗网络的用电负荷数据脱敏方法,可以实现隐私性和可用性之间的平衡,确保即使攻击者窃取了训练数据也无法从中推断真实数据信息。首先,介绍了瑞利差分隐私RDP(Rényi Differential Privacy)的基本概念以及融合RDP保护的生成对抗网络结构;然后,分三阶段介绍数据脱敏过程,理论推导所提方法如何实现对真实数据的差分隐私,并量化总隐私预算;最后,从定性和定量的角度进行算例分析,对隐私保护处理后数据的隐私性和可用性进行验证。

1　融合差分隐私保护的生成对抗网络结构

差分隐私[7]是一种备受关注的隐私保护技术,于2006年被Dwork首次提出。本文利用具有更严格隐私约束的RDP跟踪训练过程中花费的隐私预算,进而评估并最小化隐私损失,从而保护训练数据的隐私性。

1.1　RDP的基本概念

差分隐私通过对查询或分析结果添加噪声信号,在保留数据集统计学特征的前提下去除个体特征以保护个体隐私,从而使该数据库的计算处理结果对于某个具体记录(如1条负荷数据)的变化不敏感。差分隐私应用最广泛的定义为(ε,δ)-差分隐私,具体见附录A定义A1和定义A2。其中,(ε,δ)为差分隐私保护算法的隐私预算参数,表示隐私的置信水平。

将(ε,δ)-差分隐私定义应用于神经网络训练时存在如下问题:由于深度学习训练是不断迭代反向传播的过程,每次迭代时应用(ε,δ)-差分隐私机制都会使隐私预算线性增大,具体见附录A引理A1。针对这一不足,有学者基于瑞利散度提出了RDP[16],具体见附录A定义A3和定义A4。

(α,ε)-RDP有以下2个重要的基本属性。

1)引理1(RDP组合定理[16])　若给定一个随机算法$A:X\to Y_1$满足(α,ε_1)-RDP,且算法$B:Y_1\times X\to Y_2$满足(α,ε_2)-RDP,则(M_1,M_2)满足$(\alpha,\varepsilon_1+\varepsilon_2)$-RDP,其中$M_1\sim A(X)$,$M_2\sim B(M_1,X)$。

2)引理2[16]　若算法$A:X\to Y$满足(α,ε)-RDP,则对于$\forall 0<\delta<1$,A也满足$\left(\varepsilon+\dfrac{\ln(1/\delta)}{\alpha-1},\delta\right)$-差分隐私。

上述2个引理构成了本文所提隐私保护方法的理论基础:引理1将总隐私成本分解为多个串行模块的隐私成本的组合,为所提模块的隐私预算量化提供了理论依据;引理2可将RDP结果转换为传统的(ε,δ)-差分隐私形式。

最常用的RDP实现方法之一是文献[17]提出的差分隐私随机梯度下降法DPSGD(Differentially Private Stochastic Gradient Descent)。该方法通过向神经网络的反向传播梯度中加入满足高斯分布的噪声实现差分隐私,其主要步骤为:①裁剪单条数据的梯度范数以限制算法对单条数据的敏感度;②对批次数据的反向传播梯度中添加满足高斯分布的噪声;③执行梯度下降优化步骤以更新网络参数。DPSGD基于Moments Accountant组合定理,可在相同的隐私预算下达到更好的模型训练效果。

引理3(Moments Accountant组合定理[17])　设DPSGD的随机选择概率为Q,训练次数为T,M_i为标准差为σ的高斯机制,那么对于存在常数c_1和c_2使得$\forall\varepsilon<c_1Q^2T$、$\forall\delta>0$的情况,若高斯机制$M_i$的标准差$\sigma$满足式(1),则$(M_1,M_2,\cdots,M_T)$满足$(\varepsilon,\delta)$-差分隐私。

$$\sigma\geqslant c_2\frac{Q\sqrt{T\ln(1/\delta)}}{\varepsilon}\tag{1}$$

1.2　RDP保护下的时序生成对抗网络结构

基于数据合成的用户用电负荷数据脱敏的研究主要存在以下2个挑战。

1)数据隐私保护:大多数已有的负荷数据合成方法都没有在保护数据隐私的前提下训练模型,不具备数据脱敏能力。

2)时间特征保留:用户用电与温湿度、光照等随时间动态变化的因素有关,同时也与用户收入、家庭住址、节假日等静态影响因素有关。已有的生成对抗网络方法还不能很好地保留用电负荷数据特有的时间相关性和静态特征。

鉴于此,本文提出了RDP保护下的时序生成对抗网络模型,如图1所示。具体而言,RDP保护下的时序生成对抗网络由编码器、解码器、生成器和判别器组成:①对于生成对抗网络的生成器和判别器,采用能够有效捕捉时间步之间关系的循环神经网络RNN(Recurrent Neural Network),挖掘用电负荷数据的动态时间特性;②将自编码器(autoencoder)与生成对抗网络相结合,利用自编码器对用电负荷数

据的静态特征进行挖掘；③引入RDP机制，保护真实数据的敏感信息。

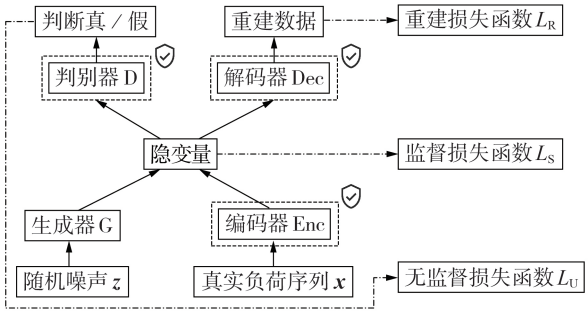


图1 RDP保护下的时序生成对抗网络模型

Fig.1 Time-series GAN model with RDP protection

自编码模块（编码器、解码器）提供负荷数据与隐变量空间（latent space）的可逆映射，从而降低对抗性学习空间的高维性，允许对抗模块（生成器、判别器）通过低维特征表示学习负荷数据的潜在影响因素。这利用了这样一个事实：复杂系统的时间特性往往也是由更少和更低维的因素驱动的。同时，引入自编码架构，有助于挖掘用户用电负荷的静态特征。

将自编码架构与对抗网络相结合，实质上是在原始生成对抗网络的无监督范式中引入训练数据特征信息进行监督训练，与原始对抗网络结构相比，利用了训练数据中更多的信息。自编码模块提供负荷数据影响因素编码的隐变量空间，对抗模块在该空间内运行，真实负荷数据与合成负荷数据的潜在影响因素通过监督损失函数进行同步，使生成器同时学习负荷数据的动态时间特性和静态特征。

上述模块设计主要关注合成负荷数据的可用性。为了实现可用性与隐私性之间的平衡，且由于编码器、解码器和判别器在训练过程中与真实用电数据有直接接触，故基于DPSGD对这3个模块添加差分隐私保护机制，实现模块参数对输入数据的差分隐私保护。生成器则是通过与编码器、解码器和判别器之间的交互进行学习训练，根据差分隐私的后处理免疫性[18]，生成器参数及其合成数据也获得同样的差分隐私保护属性。将满足差分隐私保护的合成负荷数据替代真实负荷数据用于云计算，即使合成负荷数据被攻击者非法窃取，攻击者也无法通过合成数据分辨真实数据的敏感信息，从而实现对真实数据的脱敏。

定理1 对于任意2个相邻数据集X和X'，设从负荷数据集X到生成器参数Φ的映射$A:X\rightarrow\Phi$满足(ε,δ)-差分隐私，给定生成器参数Φ到生成器输出Y的随机映射$f:\Phi\rightarrow Y$，则$f\circ A:X\rightarrow Y$满足(ε,δ)-差分隐私性质，其中“$\circ$”为复合函数运算符。具体证明过程见附录B。

2 基于生成对抗网络的用电负荷数据脱敏过程

生成对抗网络往往因其难以训练备受诟病，尤其是在差分隐私保护下对指导生成器训练的其他模块引入了噪声干扰，使生成器的训练愈加困难。在整个网络联合训练之前，先对模块进行分块预训练，可以降低生成器的训练难度，加快模型的收敛速度，提高模型训练的稳定性，在相同的隐私预算下生成更高质量的合成数据。本节介绍所提模型的训练过程，理论推导所提模型如何实现对真实用电负荷数据的差分隐私，并量化总隐私预算。基于生成对抗网络的用电负荷数据脱敏过程如附录C图C1所示，主要分为自编码模块预训练、有监督生成器预训练、联合训练3个阶段。

2.1 自编码模块预训练

自编码器是一种神经网络架构，由编码器Enc：$\mathbf{R}^n\rightarrow\mathbf{R}^d$和解码器Dec：$\mathbf{R}^d\rightarrow\mathbf{R}^n$组成，其中$n$为真实日负荷序列长度，$d$为隐变量长度。将输入的真实负荷序列$\boldsymbol{x}\in\mathbf{R}^n$映射到隐变量空间$\boldsymbol{L}\in\mathbf{R}^d$进而重构数据$\hat{\boldsymbol{x}}\in\mathbf{R}^n$。在理想的情况下，可以实现对原始输入真实负荷序列的完美重构，即$\boldsymbol{x}=\hat{\boldsymbol{x}}$。对于连续输入的数据而言，自编码器可使用均方误差MSE（Mean Square Error）作为重建损失函数L_R，如式（2）所示。

$$L_R(\boldsymbol{x},\hat{\boldsymbol{x}})=E\left[\left\|\boldsymbol{x}-\hat{\boldsymbol{x}}\right\|_2\right] \tag{2}$$

式中：$E[\cdot]$为随机变量期望。编码器Enc和解码器Dec均采用循环神经网络，以捕获负荷序列时间步之间的时间相关性。在自编码模块的训练过程中，需要访问真实负荷数据以捕获负荷的潜在特征，这样的操作可能会触及用户隐私。因此，基于DPSGD对编码器和解码器采取裁剪梯度、添加噪声的措施，可以保证真实数据的差分隐私，具体算法见附录D算法D1。

2.2 有监督生成器预训练

经过自编码模块预训练后，自编码模块可以初步提供负荷数据的隐变量特征空间。然后，利用编码器输出的负荷隐变量对生成器进行监督训练，明确鼓励模型捕获负荷的动/静态时间特征。

以监督的方式训练生成器：编码器Enc输入真实负荷序列$\boldsymbol{x}\in\mathbf{R}^n$，将其映射到隐变量空间$\boldsymbol{L}\in\mathbf{R}^d$；生成器G输入随机噪声$\boldsymbol{z}\in\mathbf{R}^n$，同样将其映射到隐变量空间$\boldsymbol{L}'\in\mathbf{R}^d$。有监督训练的损失由这2个分布之间的差异产生，使生成器学习到负荷潜在特征。应用最大似然作为监督训练的损失函数L_S，如式（3）所示。

$$L_S(\boldsymbol{x},\boldsymbol{z})=E\left[\left\|\mathrm{Enc}(\boldsymbol{x})-G(\boldsymbol{z})\right\|_2\right] \tag{3}$$

式中：$\mathrm{Enc}(\cdot)$为编码器函数；$G(\boldsymbol{z})$为生成器生成的隐变量。

生成器G采用循环神经网络，以捕获负荷序列时间步之间的相关性。在该阶段的训练中，生成器需要编码器对真实负荷数据的访问进行监督训练，因此同样需要在编码器的训练过程中加入噪声。具体算法见附录D算法D2。

2.3　联合训练

对编码器、解码器、生成器、判别器这4个模块进行联合训练。所提时序生成对抗网络模型结合了无监督范式的灵活性和监督训练具备的高效指导，通过监督目标和无监督对抗目标共同优化的负荷数据隐变量空间，使生成对抗网络在采样过程中遵循真实负荷训练数据的动/静态特征软约束。

生成对抗网络能够无监督学习训练数据的特征，拟合真实数据的分布，生成高质量的合成数据。生成对抗网络的学习过程是基于一个生成器G和一个判别器D在如下零和极小极大（即对抗性）博弈中进行的：

$$L_{\mathrm{U}}(\boldsymbol{x},\boldsymbol{z})=\min_{G}\max_{D}\left\{E\left[\log_2 D(\boldsymbol{x})\right]_{\boldsymbol{x}\sim p_{\mathrm{data}}(\boldsymbol{x})}+E\left[\log_2(1-D(G(\boldsymbol{z})))\right]_{\boldsymbol{z}\sim p(\boldsymbol{z})}\right\} \tag{4}$$

式中：L_{U}为无监督损失函数；$p_{\mathrm{data}}(\boldsymbol{x})$为真实负荷序列$\boldsymbol{x}$的分布；$p(\boldsymbol{z})$为随机噪声$\boldsymbol{z}$的先验分布；$D(\cdot)$为输出区间为[0,1]的判别器函数，采用循环神经网络。生成对抗网络的训练过程就是生成器和判别器不断博弈的过程，最终使生成器生成的样本无限接近真实数据样本，模型收敛则训练结束，达到纳什平衡。

在联合训练中：首先，作为样本空间和隐变量空间之间的可逆映射，编码器和解码器应该能够实现负荷数据的准确重建，因此第1个损失函数为重建损失函数L_{R}，见式(2)；然后，生成器和判别器之间存在无监督博弈，判别器致力于提高对真/假输入数据的判别精度，生成器希望能够生成使判别器误认为是真实数据的合成数据，因此第2个损失函数为无监督损失函数L_{U}，见式(4)；最后，为了使生成器学习到自编码模块形成的负荷特征隐变量空间，对生成器和编码器之间施加式(3)所示监督损失函数L_{S}，使生成器学习到真实数据的动/静态特征。在训练过程中，对编码器、解码器和判别器施加差分隐私随机梯度下降机制，使生成器在学习到用电负荷数据特征的同时，避免泄露真实信息。因此，自编码模块的损失函数L_{ae}、生成器的损失函数L_{g}、判别器的损失函数L_{d}分别为：

$$L_{\mathrm{ae}}=\lambda_1 L_{\mathrm{S}}+L_{\mathrm{R}} \tag{5}$$

$$L_{\mathrm{g}}=\lambda_2 L_{\mathrm{S}}+L_{\mathrm{U}} \tag{6}$$

$$L_{\mathrm{d}}=-L_{\mathrm{U}} \tag{7}$$

式中：λ_1、λ_2为比例系数。联合训练的具体算法见附录D算法D3。

2.4　总隐私预算

如RDP组合定理（引理1）所述，RDP的可组合性允许对总隐私预算的计算执行累加过程。在定理1的基础上可以得到以下推论。

推论1　在联合训练阶段，设每一个训练步中自编码模块和判别器的隐私预算分别为ε_1、ε_2，则对于任意给定的α，该训练步的总隐私预算$\varepsilon_{\mathrm{step}}=\varepsilon_1+\varepsilon_2$。

推论2　在任一训练阶段（自编码模块预训练阶段、有监督生成器预训练阶段、联合训练阶段），设每一个训练步的隐私预算为$\varepsilon_{\mathrm{step}}$，该阶段共训练$N$次，则对于任意给定的$\alpha$，该阶段的总隐私预算$\varepsilon_{\mathrm{total}}=N\varepsilon_{\mathrm{step}}$。

推论3　对于任意给定的α，设自编码模块预训练阶段、有监督生成器预训练阶段、联合训练阶段的总隐私预算分别为$\varepsilon_{\mathrm{ae}}$、$\varepsilon_{\mathrm{g}}$、$\varepsilon_{\mathrm{gan}}$，则整个训练过程的总隐私预算$\varepsilon=\varepsilon_{\mathrm{ae}}+\varepsilon_{\mathrm{g}}+\varepsilon_{\mathrm{gan}}$。

在2.1—2.3节中，基于DPSGD在每一个训练步实现对输入数据集的RDP。如RDP组合定理（引理1）所述，对于任意给定的α，当多个顺序组合的RDP机制作用于同一个数据集X时，总隐私预算存在可累加性。值得注意的是，此处的多个顺序组合的差分隐私机制并不要求是相同的算法，而是强调对同一个数据集的多次访问。它允许多个不同的数据访问方式，共同消耗有限的隐私预算。因此，在联合训练阶段，单个训练步中先后出现2次用电负荷数据访问（分别作为编码器和判别器的输入），此时单步总隐私预算为2次数据访问所消耗的隐私预算之和（推论1）；任意训练阶段的总隐私预算为各个训练步所消耗的隐私预算之和（推论2）；整个训练过程的总隐私预算为3个阶段所消耗的隐私预算之和（推论3）。

在根据所添加的高斯噪声分布计算总隐私预算ε时，采用数值积分方法[17]可以使计算结果更加准确。为了得到一个更严格的隐私上界，对多个α取值进行计算。取其中最小的ε及其对应的α作为最终的(α,ε)-RDP隐私保护水平。根据引理2，可将其转换为更普遍的(ε,δ)-差分隐私形式。

3　算例分析

基于浙江省某地区的真实用电负荷数据集进行算例验证与分析。该数据集包括2019年7月1日至8月31日共885条日负荷数据，采样频率为15 min/次。在配置NVIDIA Quadro RTX 4000的工作站中采用图形处理器（GPU）计算模式，编程环境为Python 3.6/PyTorch[19]。

本文所提用电负荷数据脱敏方法只在本地（终端）部署和运行，算法本身不存在被攻击的途径，因此只对合成用电负荷数据的质量进行分析评价。考

虑到负荷数据大多是无标签的，提出以下3个评价指标对合成用电负荷数据的质量进行分析：①隐私性，从合成数据中无法推断真实数据信息；②真实性，合成数据应当与真实数据具有相同的分布；③可用性，当应用于同一目的（如预测、分类等）时，合成数据应当与真实数据具有相同的效果。三者之间的关系为：合成负荷数据的隐私性要求从合成数据中无法推断出使单条真实负荷数据的敏感信息，且不希望改变负荷数据集的整体分布特性（真实性）和负荷数据自身的特性（可用性）。

3.1 合成负荷数据的隐私性验证

合成负荷数据的隐私性要求即使攻击者获得了合成数据，也无法从中获得真实数据信息。采取基于相似度的指标来衡量攻击者从合成数据中识别得到个人真实数据信息的可能性。

对合成负荷数据集随机采样885条数据，计算每条合成数据与原始数据中最接近的记录之间的欧氏距离d_o，如式（8）所示。

$$d_o=\sqrt{\sum_{i=1}^{n}(x_i-y_i)^2} \tag{8}$$

式中：x_i、y_i分别为任意1条真实负荷数据和合成负荷数据中采样点i的量测值。当d_o小于某一相似度阈值时，可以认为合成数据与真实数据相匹配；当d_o=0时，表明合成数据与某条真实数据相吻合，即泄露了真实信息。将可以被合成数据匹配到的真实数据占全部真实数据的比例定义为匹配率。

不同相似度阈值下合成数据与真实数据的匹配率如图2所示。以真实负荷数据间的最小欧氏距离作为参考，由于各条负荷数据与其他负荷数据间的最小欧氏距离的平均值为0.5，故在小于0.5的范围内，取0.30、0.35、0.40作为相似度阈值。由图2可知：当相似度阈值为0.40时，随着ε增大，匹配率随之增大，最终稳定在22%左右，即大约有22%的真实数据与合成数据之间的最小欧氏距离小于0.40，有一定的可能被合成数据泄露部分信息；当相似度阈值为0.35时，仅有极少量的真实数据可以与合成数据相匹配；而当相似度阈值为0.30时，没有真实数据可以与合成数据相匹配，即在任意的总隐私预算下，合成数据与真实数据之间的最小欧氏距离均没有小于0.30的情况，说明本文所提方法合成的负荷数据具有较好的隐私性。

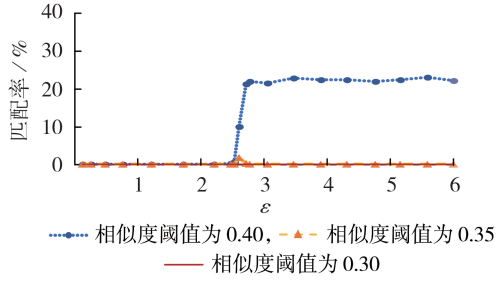


图2 不同相似度阈值下的匹配率

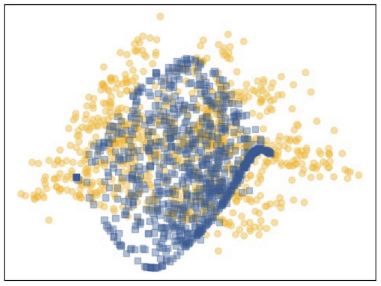
Fig.2 Matching rate under different similarity thresholds

3.2 合成负荷数据的真实性验证

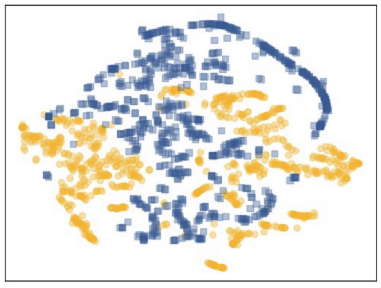
本节从定性和定量2个角度比较合成负荷数据与真实负荷数据的分布。

1）分布可视化。对合成数据集随机采样885条数据，利用t分布随机邻接嵌入（t-SNE）[20]算法和主成分分析（PCA）[21]算法对合成样本与原始样本在二维空间中的分布进行可视化，从而对合成负荷数据的真实性进行定性评估。

当总隐私预算ε=5时，PCA算法和t-SNE算法的降维可视化结果如图3所示，当总隐私预算ε分别为1和无穷大（即不设隐私保护机制）时的可视化结果对比如附录E图E1所示。由图可以看出，随着总隐私预算的增大，合成负荷数据分布逐渐趋近真实负荷数据分布，在较小的总隐私预算（ε=5）下合成数据已经具有与真实数据相似的分布，且在不设置隐私保护机制（ε为无穷大）时合成负荷数据分布与真实负荷数据分布基本吻合，表明所提方法有效地捕捉了负荷特征和时间相关性。



（a）PCA算法



（b）t-SNE算法

真实负荷数据，合成负荷数据

图3 ε=5时PCA算法和t-SNE算法的降维可视化结果

Fig.3 Visualized results of dimension reduction for PCA algorithm and t-SNE algorithm when ε=5

2）最大均值差异MMD（Maximum Mean Discrepancy）。最大均值差异是一种用于度量不同域数据集之间分布差异的指标，可以用来表示模型捕获

真实数据分布的程度，其定义如下：

$$\gamma_{\mathrm{MMD}}[F,p,q]=\sup_{\|f\|_H\leqslant 1} E_p[f(\boldsymbol{x})]-E_q[f(\boldsymbol{y})] \quad (9)$$

式中：F为再生核希尔伯特空间H中的单位球；p、q分别为随机变量$\boldsymbol{x}$、$\boldsymbol{y}$的分布；sup表示上确界；$f(\cdot)$为映射函数，且$f\in F$；$\|f\|_H$为函数f在再生核希尔伯特空间H中的范数。最大均值差异γ_{MMD}越小，表明2个数据集之间的分布相似性越高。对合成数据随机采样885条数据，计算合成数据与真实数据间的最大均值差异，从而对合成负荷数据的真实性进行定量评估。

最大均值差异与总隐私预算之间的关系曲线见图4。由图可知，在差分隐私保护下，当隐私预算较小时，随着ε增大，合成数据与真实数据之间的分布差异逐渐减小；当隐私预算增大到一定的阈值后，二者之间的最大均值差异逐渐收敛到一个较小的定值，并逐渐趋近于无隐私保护时的数值。图4验证了所提方法对真实负荷分布的捕获能力，合成负荷数据的分布与真实数据分布之间差别较小，从定量角度验证了合成负荷数据的真实性。

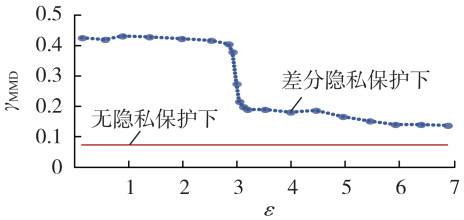


图4　最大均值差异与总隐私预算之间的关系曲线

Fig.4　Relationship curves between MMD and ε

3.3　合成负荷数据的可用性验证

设定如下MLaaS应用场景：用户将合成负荷数据作为训练集上传至云平台，利用云平台提供的预测模型进行负荷预测。设云平台采用目前在时间序列预测领域得到最广泛应用的长短期记忆LSTM（Long-Short Term Memory）神经网络[22]作为负荷预测模型。合成负荷数据的可用性意味着合成数据应当与真实数据具有相同的预测效果。

引入3组不同的训练集、测试集进行实验：①实验1，基于真实数据集训练预测模型，基于真实测试集测试模型的性能；②实验2，对合成训练集进行训练，对真实测试集进行测试；③实验3，对合成训练集进行训练，对合成测试集进行测试。其中，训练集和测试集在真实数据集和合成数据集中都是不相交的。一方面，如果基于合成负荷数据训练预测模型，应用于真实数据时有较高的预测性能（实验2），则合成负荷数据很好地捕捉了真实负荷数据的时序特征。同时，这也意味着利用经过隐私保护处理的数据代替用户敏感数据用于模型训练是可行的。另一方面，如果基于合成数据对负荷预测模型进行训练和测试（实验3）时的性能与基于真实数据进行训练和测试（实验1）时的性能相似，则允许研究人员将大量的实验部署在基于合成数据进行算法调试，只需要对真实敏感数据进行少量的测试，从而降低隐私保护成本。

采用平均绝对误差MAE（Mean Absolute Error）计算预测误差，作为合成负荷数据可用性的定量评估。3组实验的负荷预测平均绝对误差（标幺值）如图5和表1所示。由结果可知，当将基于合成负荷数据训练的预测模型应用于真实数据（实验2）时，其预测误差随着ε的增大而减小，当ε超过一定的阈值时，其预测误差近似等于基于真实负荷数据训练的预测模型（实验1）的结果，甚至有偏小情况出现。这说明本文所提方法很好地学习到了负荷数据的时序特征，即使没有学习到真实负荷数据中较为敏感的部分信息，也可以将预测误差控制在较小的范围内。相较于基于真实数据训练和测试负荷预测模型（实验1）的性能，基于合成数据训练和测试负荷预测模型（实验3）时，当ε较小时，预测误差也较小，表明此时的合成数据质量较低，时序特性较差；随着ε增大，预测误差逐渐趋近于实验1的预测误差，表明此时基于合成数据进行预测已经具有与真实数据相似的性能。上述3组实验结果通过定量的方式验证了合成负荷数据的可用性。

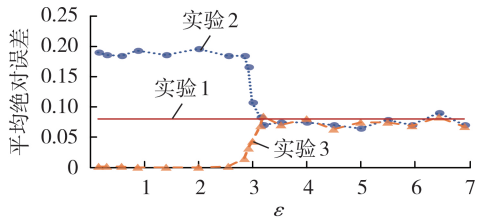


图5　负荷预测平均绝对误差与总隐私预算的关系曲线

Fig.5　Relationship curves between MAE of load prediction and ε

表1　负荷预测平均绝对误差

Table 1　MAE of load prediction

ε	平均绝对误差		
	实验1	实验2	实验3
0.1	0.079	0.189	0.002
1	0.079	0.186	0.001
3	0.079	0.107	0.043
5	0.079	0.064	0.074
7	0.079	0.070	0.068

综合3组实验的结果可知，负荷数据的隐私性与其真实性、可用性之间存在负相关关系，增大隐私保护力度势必会导致捕获到的负荷特征信息减少。因此在实际应用中，需要结合具体的需求选取合适

的隐私预算:若实际应用场景中对算法精度、数据可用性要求较高,则可选择较大的隐私保护预算;若实际应用场景对数据隐私要求较高而可以允许牺牲一部分算法精度和数据可用性,则应选择较小的隐私预算。

4 结论

本文针对云平台MLaaS中可能产生的敏感数据泄露问题,通过构建融合差分隐私机制、自编码器、生成对抗网络的用电负荷数据生成模型,以采用满足差分隐私的合成数据替代真实数据实现数据脱敏。通过算例进行实验分析,可得如下结论:

1)所提基于时序生成对抗网络的用电负荷数据生成模型,实现了对用户用电隐私数据的差分隐私保护及隐私预算的量化;

2)在一定的隐私预算下,本文所提方法能够生成分布与真实负荷数据分布相近、可用性与真实负荷数据相似的合成负荷数据,在有效保护用户用电隐私的同时,保证了合成数据的可用性。

训练数据隐私泄露已成为阻碍云计算进一步发展的重要原因之一。如何在保证合成数据可用性的前提下尽可能减少真实数据的隐私保护开销,是需要进一步解决的问题。随着新型电力系统的发展,电网运营商对云端计算资源的需求将进一步加大,解决云计算模式下的隐私保护问题刻不容缓,本文所提方法和思路具有较好的应用前景。

附录见本刊网络版(http://www.epae.cn)。

参考文献:

[1] 张亚健,杨挺,孟广雨. 泛在电力物联网在智能配电系统应用综述及展望[J]. 电力建设,2019,40(6):1-12.
ZHANG Yajian,YANG Ting,MENG Guangyu. Review and prospect of Ubiquitous Power Internet of Things in smart distribution system[J]. Electric Power Construction,2019,40(6):1-12.

[2] 国家发展和改革委员会. "东数西算"工程系列解读之三"东数西算"助力中国数字经济均衡发展[EB/OL]. [2022-03-10]. https://www.ndrc.gov.cn/xxgk/jd/jd/202203/t20220317_1319465.html.

[3] 孙宇嫣,蔡泽祥,马国龙,等. 电力物联网云主站计算负荷模型与资源优化配置[J]. 电力自动化设备,2021,41(4):177-183.
SUN Yuyan,CAI Zexiang,MA Guolong,et al. Workload model and optimal resource allocation of cloud master station in Power Internet of Things[J]. Electric Power Automation Equipment,2021,41(4):177-183.

[4] 谭作文,张连福. 机器学习隐私保护研究综述[J]. 软件学报,2020,31(7):2127-2156.
TAN Zuowen,ZHANG Lianfu. Survey on privacy preserving techniques for machine learning[J]. Journal of Software,2020,31(7):2127-2156.

[5] 郭红霞,陆进威,杨苹,等. 非侵入式负荷监测关键技术问题研究综述[J]. 电力自动化设备,2021,41(1):135-146.
GUO Hongxia,LU Jinwei,YANG Ping,et al. Review on key techniques of non-intrusive load monitoring[J]. Electric Power Automation Equipment,2021,41(1):135-146.

[6] 陈中,方国权,赵奇,等. 面向区域级用户的非侵入式负荷监测技术应用方法[J]. 电力自动化设备,2020,40(8):126-132.
CHEN Zhong,FANG Guoquan,ZHAO Qi,et al. Application method of non-intrusive load monitoring technology for region-level users[J]. Electric Power Automation Equipment,2020,40(8):126-132.

[7] DWORK C,ROTH A. The algorithmic foundations of differential privacy[J]. Foundations and Trends in Theoretical Computer Science,2014,9(3/4):211-407.

[8] EIBL G,ENGEL D. Differential privacy for real smart metering data[J]. Computer Science-Research and Development,2017,32(1/2):173-182.

[9] TORFI A,FOX E A,REDDY C K. Differentially private synthetic medical data generation using convolutional GANs[J]. Information Sciences,2022,586:485-500.

[10] GOODFELLOW I J,POUGET-ABADIE J,MIRZA M,et al. Generative adversarial networks[J]. Advances in Neural Information Processing Systems,2014,3:2672-2680.

[11] 李元诚,杨珊珊. 基于改进自注意力机制生成对抗网络的智能电网GPS欺骗攻击防御方法[J]. 电力自动化设备,2021,41(11):100-106.
LI Yuancheng,YANG Shanshan. Defense method of smart grid GPS spoofing attack based on improved self-attention generative adversarial network[J]. Electric Power Automation Equipment,2021,41(11):100-106.

[12] 张承圣,邵振国,陈飞雄,等. 基于条件深度卷积生成对抗网络的新能源发电场景数据迁移方法[J/OL]. 电网技术. (2021-07-23)[2022-03-10]. https://doi.org/10.13335/j.1000-3673.pst.2021.1008.

[13] 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):56-64,320.
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):56-64,320.

[14] 殷豪,张铮,丁伟锋,等. 基于生成对抗网络和LSTM-CSO的少样本光伏功率短期预测[J/OL]. 高电压技术. (2021-09-22)[2022-03-10]. https://doi.org/10.13336/j.1003-6520.hve.20210946.

[15] HITAJ B,ATENIESE G,PEREZ-CRUZ F. Deep models under the GAN:information leakage from collaborative deep learning[C]//Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security. Dallas,Texas,USA:ACM,2017:603-618.

[16] MIRONOV I. Rényi differential privacy[C]//2017 IEEE 30th Computer Security Foundations Symposium. Santa Barbara,CA,USA:IEEE,2017:263-275.

[17] ABADI M,CHU A,GOODFELLOW I,et al. Deep learning with differential privacy[C]//Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security. Vienna,Austria:ACM,2016:308-318.

[18] XIE L,LIN K,WANG S,et al. Differentially private generative adversarial network[EB/OL]. (2018-02-19)[2022-03-10]. https://doi.org/10.48550/arXiv.1802.06739.

[19] PASZKE A,GROSS S,MASSA F,et al. PyTorch:an imperative style,high-performance deep learning library[C]//Proceedings of the 33rd International Conference on Neural Information Processing Systems. Red Hook,NY,USA:[s.n.],2019:8026-

8037.
[20] LAURENS V D M，HINTON G. Visualizing data using t-SNE [J]. Journal of Machine Learning Research，2008，2605(9)：2579-2605.
[21] BRYANT F B，YARNOLD P R. Principal-components analysis and exploratory and confirmatory factor analysis[M]. Washington DC，USA：American Psychological Association，1995：99-136.
[22] 蔡昌春，息梦蕊，刘昊林，等. 基于数据驱动和多场景技术的微电网并网等效建模[J/OL]. 电力自动化设备. [2022-03-10]. https://doi.org/10.16081/j.epae.202203009.

作者简介：

于　群

于　群(1998—)，女，博士研究生，主要研究方向为机器学习、能源大数据分析等(**E-mail**：yuqun@zju.edu.cn)；

李知艺(1989—)，男，研究员，博士研究生导师，博士，通信作者，主要研究方向为智能配电网、能源大数据分析等(**E-mail**：zhiyi@zju.edu.cn)。

(**编辑**　陆丹)

Differential privacy protection method of electrical load data towards cloud computing applications

YU Qun[1]，SHEN Zhiheng[2]，SUN Feifei[2]，LI Zhiyi[1]

(1. College of Electrical Engineering，Zhejiang University，Hangzhou 310027，China；
2. Economic and Technology Research Institute of State Grid Zhejiang Electric Power Company，Hangzhou 310016，China)

Abstract：With the development of cloud computing technology，users can use public computing resources to complete big data analysis services such as machine learning with low cost and high efficiency，which improves computing efficiency and economic benefits while facing privacy disclosure risk. Aiming at the users' load data leakage problem hidden in the cloud computing represented by machine learning as a service，an electrical load data masking method based on time-series generative adversarial network under differential privacy protection is proposed. The synthetic data satisfying differential privacy is used to replace the original sensitive data so as to effectively prevent attackers from inferring real sensitive information from stolen training data. Rényi differential privacy mechanism is introduced to remove individual characteristics on the premise of keeping the statistical characteristics of load data. On this basis，the recurrent neural network is used as generator and discriminator of generative adversarial network to capture the dynamic time characteristics of load series. At the same time，the static characteristics of load series are mined by combining autoencoder with generative adversarial network. Theoretical derivation proves that the proposed method can meet the differential privacy requirements and the total privacy budget can be quantified. Numerical experiment results verify that the proposed method can ensure the privacy and availability of electrical load data after privacy protection processing.

Key words：electrical load data；cloud computing；differential privacy protection；generative adversarial networks；autoencoders；data masking

附录 A

定义 A1（相邻数据集[7]） 如果 2 个结构和属性相同的数据集 D 和 D'，它们除了某一条数据不同外，其余数据都相同，那么可以称数据集 D 和 D' 为相邻数据集。

定义 A2（(ε,δ)-差分隐私[7]） 对于任意 2 个相邻数据集 D 和 D'，给定一个随机算法 $A:D\to Y$，如果 A 任意的查询结果 Y 满足式(B1)，则随机算法 A 满足 (ε,δ)-差分隐私。

$$P[A(D)\in Y]\le \mathrm{e}^{\varepsilon}P[A(D')\in Y]+\delta \tag{A1}$$

式中：参数 (ε,δ) 为差分隐私保护算法的隐私预算参数，表示隐私的置信水平。(ε,δ) 越小，差分隐私保护水平越高。

引理 A1（组合定理[7]） 若给定一个随机算法 M_i 满足 (ε_i,δ_i)-差分隐私，那么 $(M_1,M_2,\cdots,M_k)$ 满足 $\left(\sum_{i=1}^{k}\varepsilon_i,\sum_{i=1}^{k}\delta_i\right)$-差分隐私。

定义 A3（α阶瑞利散度[16]） 分布 P 与分布 P' 的α阶瑞利散度定义为：

$$D_{\alpha}(P\,\|\,P')=\frac{1}{\alpha-1}\ln\left(E_{P'(X)}\left[\frac{P(X)}{P'(X)}\right]^{\alpha}\right) \tag{A2}$$

定义 A4（(α,ε)-RDP[16]） 对于任意 2 个相邻数据集 D 和 D'，给定一个随机算法 $A:D\to Y$，如果 A 任意的查询结果 Y 满足式(A3)，则随机算法 A 满足 (α,ε)-RDP。

$$D_{\alpha}(A(D)\,\|\,A(D'))\le\varepsilon \tag{A3}$$

附录 B

定理 1 对于任意 2 个相邻数据集 D 和 D'，设从负荷数据 D 到生成器参数Φ的映射 $A:D\to\Phi$是满足 (ε,δ)-差分隐私的，给定生成器参数Φ到生成器输出 Y 的随机映射 $f:\Phi\to Y$，则 $f\circ A:D\to Y$ 是 (ε,δ)-差分隐私的。

证明：首先证明在给定生成器输入噪声 z 的情况下，上述命题成立。不失一般性地，设生成器为一个全连接网络。设生成器除输入层外共有 H 层，第 l（l=1，2，…，H）层的网络权重为 $W^{(l)}$，网络偏置为 $b^{(l)}$，输出为 $x^{(l)}$，激活函数为$\sigma^{(l)}$。则：

$$x^{(l)}=\begin{cases}\sigma^{(l)}(W^{(l)}x^{(l-1)}+b^{(l)}) & 1\le l\le H\\ z & l=0\end{cases} \tag{B1}$$

此时生成器参数Φ=$\{W, b\}$到生成器输出 Y=$x^{(H)}$的映射 f: $\Phi\to Y$ 为一确定性函数。则对于任意的相邻负荷数据集 D 和 D'，给定任意生成器输出 $S\subseteq Y$，使 T=$\{\varphi\in\Phi: f(\varphi)\ \in Y\}$，有：

$$P[f(A(D))\in S]=P[A(D)\in T]\le e^{\varepsilon}P[A(D')\in T]+\delta=\mathrm{e}^{\varepsilon}P[f(A(D'))\in S]+\delta \tag{B2}$$

当输入噪声 z 为满足一定先验分布的随机变量时，$f:\Phi\to Y$ 为一随机映射，可以分解为确定性函数的凸组合，由于差分隐私的凸组合是差分隐私的，证明完毕。

附录 C

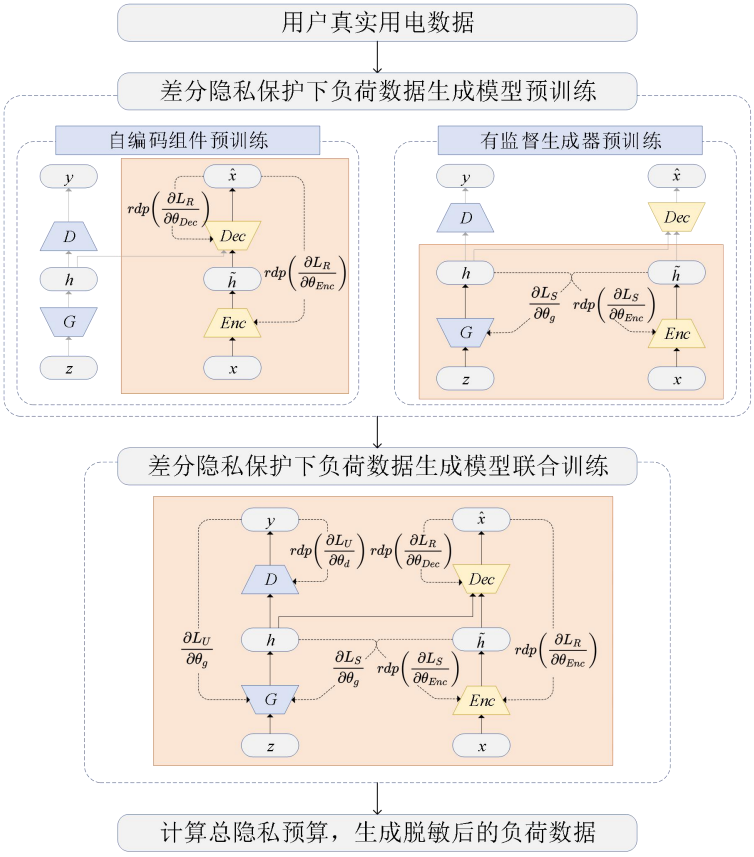


图 C1 基于生成对抗网络的负荷数据脱敏过程

Fig.C1 Load data masking process based on GAN

附录 D

算法 D1：差分隐私自编码模块预训练算法

输入：真实用户用电数据 $X=\{x_i\}_{i=1}^{N}$，学习率η，自编码模块网络参数θ，自编码模块训练次数 n_{ae}，微批次大小 k，范数上限 C，高斯噪声标准差σ

for i=1…n_{ae} **do**

采样 n 个真实样本 $X=\{x_i\}_{i=1}^{n}$ 作为一个批次

将批次 X 分为微批次 $X_1,...,X_r$，其中 $r=\left\lceil \frac{n}{k} \right\rceil$

for l=1…r **do**

$L=MSE(X_l,\hat{X}_l)$，$\hat{X}_l=Dec(Enc(X_l))$

$g_{\theta,l}\leftarrow\nabla_\theta L(\theta,X_l)$

$\hat{g}_{\theta,l}\leftarrow g_{\theta,l}/\max(1,\frac{\left\|g_{\theta,l}\right\|_2}{C})$

end for

$\hat{g}_\theta\leftarrow\frac{1}{r}\sum_{l=1}^{r}(\hat{g}_{\theta,l}+N(0,\sigma^2C^2))$

更新参数： $\hat{\theta}\leftarrow\theta-\eta\hat{g}_\theta$

end for

算法 D2：有监督生成器预训练算法

输入：真实用户用电数据 $X=\{x_i\}_{i=1}^{N}$，随机噪声 z~N(0,1)，学习率η，生成器网络参数φ，自编码模块网络参数θ，有监督生成器训练次数 n_{g}，微批次大小 k，范数上限 C，高斯噪声标准差σ

for i=1…n_{g} **do**

采样 n 个真实样本 $X=\{x_i\}_{i=1}^{n}$ 作为一个批次

采样 n 个噪声向量 $Z=\{z_i\}_{i=1}^{n}$ 作为一个批次

将批次 X 分为微批次 $X_1,...,X_r$

将批次 Z 分为微批次 $Z_1,...,Z_r$

for l=1…r **do**

$L=L_S(X_l,Z_l)$，$L_S(X_l,Z_l)=E\left[\left\|Enc(X_l)-G(Z_l)\right\|_2\right]$

$g_{\varphi,l}\leftarrow\nabla_\varphi L(\varphi,Z_l)$

$g_{\theta,l}\leftarrow\nabla_\theta L(\theta,X_l)$

$\hat{g}_{\theta,l}\leftarrow g_{\theta,l}/\max(1,\frac{\left\|g_{\theta,l}\right\|_2}{C})$

end for

$\hat{g}_\varphi\leftarrow\frac{1}{r}\sum_{l=1}^{r}(\hat{g}_{\varphi,l})$

$\hat{g}_\theta\leftarrow\frac{1}{r}\sum_{l=1}^{r}(\hat{g}_{\theta,l}+N(0,\sigma^2C^2))$

更新参数： $\hat{\varphi}\leftarrow\varphi-\eta\hat{g}_\varphi$，$\hat{\theta}\leftarrow\theta-\eta\hat{g}_\theta$

end for

算法 D3：联合训练算法

输入：真实用户用电数据 $X=\{x_i\}_{i=1}^{N}$，随机噪声 z~N(0,1)，学习率η，生成器网络参数φ，判别器网络参数 w，自编码模块网络参数θ，联合训练次数 n_{gan}，微批次大小 k，范数上限 C，高斯噪声标准差σ，比例系数λ_1、λ_2

for i=1…n_{gan} **do**

采样 n 个真实样本 $X=\{x_i\}_{i=1}^{n}$ 作为一个批次

采样 n 个噪声向量 $Z=\{z_i\}_{i=1}^{n}$ 作为一个批次

将批次 X 分为微批次 $X_1,...,X_r$

将批次 Z 分为微批次 $Z_1,...,Z_r$

for l=1…r **do**

$$L_{ae}=\lambda_1 L_S+L_R$$

$$g_{\theta,l}\leftarrow\nabla_\theta L_{ae}(\theta,X_l)$$

$$\hat{g}_{\theta,l}\leftarrow g_{\theta,l}/\max(1,\frac{\|g_{\theta,l}\|_2}{C})$$

$$L_g=\lambda_2 L_S+L_U$$

$$g_{\varphi,l}\leftarrow\nabla_\varphi L_g(\varphi,Z_l)$$

$$L_d=-L_U$$

$$g_{w,l}\leftarrow\nabla_w L(w,X_l)$$

$$\hat{g}_{w,l}\leftarrow g_{w,l}/\max(1,\frac{\|g_{w,l}\|_2}{C})$$

end for

$$\hat{g}_\theta\leftarrow\frac{1}{r}\sum_{l=1}^{r}(\hat{g}_{\theta,l}+N(0,\sigma^2C^2))$$

$$\hat{g}_\varphi\leftarrow\frac{1}{r}\sum_{l=1}^{r}(g_{\varphi,l})$$

$$\hat{g}_w\leftarrow\frac{1}{r}\sum_{l=1}^{r}(\hat{g}_{w,l}+N(0,\sigma^2C^2))$$

更新参数： $\hat{\theta}\leftarrow\theta-\eta\hat{g}_\theta$，$\hat{\varphi}\leftarrow\varphi-\eta\hat{g}_\varphi$，$\hat{w}\leftarrow w-\eta\hat{g}_w$

end for

附录 E

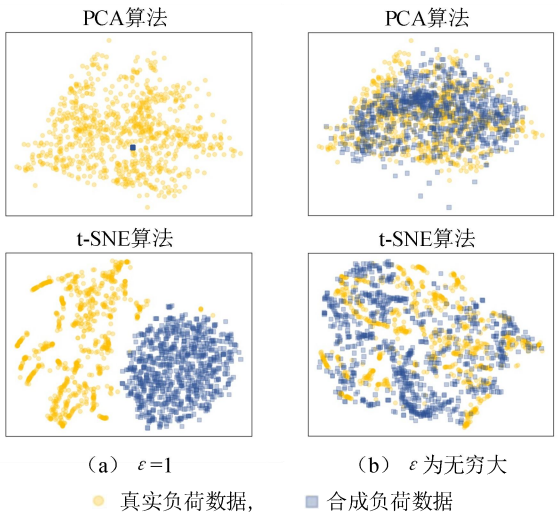


图 E1 ε 为 1 和无穷大时 PCA 算法和 t-SNE 算法的可视化结果

Fig.E1 Visualized results of PCA algorithm and t-SNE algorithm when ε is 1 and infinite