ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 36,No. 1,Jan. 2021,pp. 85－94
DOI:10. 16337/j. 1004-9037. 2021. 01. 008

http://sjcj. nuaa. edu. cn
E-mail:sjcj@nuaa. edu. cn
Tel/Fax:＋86-025-84892742

支持差分隐私的图像数据挖掘方法研究

杨云鹿[1]，周亚建[2]，宁　华[1]

(1. 中国信息通信研究院移动应用创新与治理技术工业和信息化部重点实验室,北京 100191;2. 北京邮电大学网络空间安全学院,北京 100876)

摘　要：针对数据挖掘模型中存在的隐私泄漏问题及现有隐私保护技术的不透明性,本文将差分隐私与图像生成模型生成对抗网络(Generative adversarial network, GAN)相结合,提出了一种更具普适性的支持图像数据差分隐私保护的生成对抗网络模型(Image differential privacy-GAN, IDP-GAN)。IDP-GAN通过差分隐私的拉普拉斯实现机制,将拉普拉斯噪声合理地分配到判别器的仿射变换层的输入特征以及输出层的损失函数的多项式近似系数中。在实现差分隐私保护的同时,有效地减少了训练过程中隐私预算的消耗。标准数据集MNIST和CelebA上的实验验证了IDP-GAN可以生成更高质量的图像数据,此外用成员推理攻击实验证明了IDP-GAN具有较好的抗攻击能力。

关键词：差分隐私;数据挖掘;图像生成;生成对抗网络

中图分类号：TP399　　**文献标志码**:A

Image Data Mining Method Supporting Differential Privacy

YANG Yunlu[1], ZHOU Yajian[2], NING Hua[1]

(1. Key Laboratory of Mobile Application Innovation and Governance Technology, Ministry of Industry and Information Technology, China Academy of Information and Communication Technology, Beijing 100191, China; 2. School of Cyberspace Security, Beijing University of Posts and Telecommunications, Beijing 100876, China)

Abstract: Aiming at the privacy leakage problem in the data mining model and the opacity of existing privacy protection technologies, a more universal image differential privacy-generative adversarial network (IDP-GAN) combining differential privacy with the image generation model—generative adversarial network (GAN) is proposed. IDP-GAN uses the Laplace implementation mechanism to reasonably allocate Laplace noise to the input features of the affine transformation layer and the polynomial approximation coefficients of the loss function of the output layer. While achieving differential privacy protection, IDP-GAN effectively reduces the consumption of privacy budget during training. Experiments on the standard data sets MNIST and CelebA verify that IDP-GAN can generate higher quality image data. In addition, membership inference attacks experiments prove that IDP-GAN has better ability to resist attacks.

Key words: differential privacy; data mining; image generation; generative adversarial network

收稿日期:2020-07-10;**修订日期**:2020-09-30

引　言

随着大数据与计算机视觉技术的广泛应用,语义丰富的图像背后蕴藏的巨大价值得以挖掘,同时也带来了不容小觑的隐私泄漏问题[1]。若直接发布或使用语义丰富的图像,有可能会造成严重的隐私信息泄露。因此,如何在数据挖掘的应用中实现图像数据安全性与可用性之间的平衡日益成为重要的研究方向[2-3]。

近年来,深度学习在计算机视觉、图像处理、目标检测等领域成效卓然,备受国内外研究者的关注。深度学习是用于分析大规模数据集的代表性数据挖掘方法,通常需要大量的训练样本才能达到预期的效果。但是,在某些领域中,例如医疗、军事和旅游等,无法获得足量满足需求的样本数据。针对这个问题,Goodfellow等[4]提出了一个通过对抗过程估计生成模型的深度学习框架,即生成对抗网络(Generative adversarial networks,GAN)。GAN通过学习一组训练样本中的数据分布,可以从分布中进行采样并生成更多的具有同样分布的样本,以增加原始数据集的规模。在此基础上,GAN及其变体结合了深度神经网络和博弈论的复杂性,生成了难以与原始数据区分的高质量"假"样本数据。

GAN虽然拥有出色的性能,但它也同样存在着泄漏训练样本隐私信息的风险[4]。在深度神经网络中,对抗性的训练过程和高度复杂的模型结构共同作用生成了基于训练样本的数据分布。由于深度神经网络模型的高度复杂性,GAN可以轻松地记住训练样本。如果对生成的数据分布进行反复采样,训练的原始样本有极大的概率可以被恢复。例如Hitaj等[5]提出了一种主动推理攻击模型,该模型可以从生成的"假"样本中重建训练样本。因此,GAN在隐私图像数据的应用不仅是通过使用高质量的"假"数据生成模型向公众或个人发布"假"数据,还应考虑使用隐私保护技术以减轻GAN的隐私泄漏。

为了解决上述问题,研究者们提出了多种隐私保护模型及方法[6]。Papernot等[7]提出了教师模型全体的隐私聚合(Private aggregation of teacher ensembles,PATE),该方法包含多个使用隐私数据训练的教师模型和一个基于GAN模型的学生模型,教师模型结合差分隐私对GAN评分以生成最优的学生模型。PATE仅使用学生模型进行预测,不公布教师模型,从而保护训练数据集。但是,GAN模型训练过程是不透明的,PATE隐私保护机制的可控性较弱。Abadi等[8]于2016年提出了时刻统计(Moments accountant),该方法在随机梯度下降中引入差分隐私,可以在神经网络的训练中执行适当的隐私保护。在时刻统计的基础上,Zhang等[9]提出了在数据训练阶段引入差分隐私的生成对抗网络模型(differential privacy-GAN, dp-GAN),通过将噪声注入神经网络的权重以缓解GAN的信息泄漏。然而,该模型在训练中每次引入参数的"梯度"中的噪声大小都和隐私预算与训练时期的数量成正比。因此,在实际应用中,模型可能会消耗大部分不必要的隐私预算,以这种方式将差分隐私与GAN结合会对训练GAN模型的稳定性和可伸缩性产生重大影响。此外,由于数据的语义丰富,并且缺乏有关分析任务的先验知识,引入过多的噪声以确保隐私,导致了生成数据的可用性被大幅度减弱。因此,数据挖掘领域亟需一种既可以保护隐私信息又可以产生高质量图像的生成模型。

鉴于此,本文针对生成对抗网络隐私保护过程的不透明性以及隐私预算消耗量过大的问题,提出了一种支持图像数据差分隐私保护的生成对抗网络模型(Image differential privacy-GAN, IDP-GAN),采用在训练过程中引入差分隐私的方法,将拉普拉斯噪声合理地分配到判别器的放射变层和损失函数中,以提高GAN的隐私性。更重要的是,不同于dp-GAN,IDP-GAN在训练过程中仅单次引入噪声,有效减少了隐私预算的消耗,在实现差分隐私保护的同时,提高了生成图像的使用精度。

本文的主要贡献如下:

(1)结合差分隐私技术与生成对抗网络模型,提出一种更具普适性的支持图像隐私保护的数据挖掘模型。模型在保护原始数据隐私的同时,不仅可以产生高质量生成图像,而且具有较好的抗攻击能力。

(2)模型在判别器的仿射变换层和损失函数中均以单次引入拉普拉斯噪声的方式实现差分隐私,能够减少模型训练过程中隐私预算的消耗,显著提高了模型对于大规模图像数据集的可用性。

(3)不同于将模型训练过程视作黑盒的隐私保护机制,IDP-GAN的隐私保护机制是透明且可控的,根据不同复杂度的数据集,合理分配隐私预算,可以生成更清晰、质量更好的图像。

1 相关技术概述

本节主要介绍差分隐私和生成对抗网络模型的背景知识,以及用于隐私预算分配的层级相关性传播方法。

1.1 差分隐私

差分隐私是由Dwork等[10]提出的基于严格数学理论的隐私保护模型。差分隐私主要通过添加噪声来对原始数据的变换或统计结果进行扰动,从而达到保护隐私的效果,且不会影响整体数据的分析结果。差分隐私技术的主要优点在于其具有较高的普适性及可解释性。

1.1.1 差分隐私定义

定义1[11] 存在两个相邻数据集D和D',两者的区别在于至多只有一条数据不同,用Range(M)表示随机算法M的取值范围,用$P_r[E]$表示事件E的泄露风险,若随机算法M在数据集上的取值结果$S\in$Range(M),满足不等式

$$P_r[M(D)\in S]\leqslant e^{\varepsilon}P_r[M(D')\in S] \tag{1}$$

则算法M满足ε-差分隐私。参数ε表示隐私保护预算,其值越小隐私保护程度越高。

1.1.2 差分隐私实现机制

差分隐私的常用实现机制[11]包括拉普拉斯机制和指数机制。根据图像数据的特性,本文选取拉普拉斯机制在生成对抗网络模型上实现差分隐私。

(1) 拉普拉斯机制

给定一个映射函数$f:D^n\to R^d$,表示数据集D到一个d维空间的映射关系,并且$\varepsilon>0$。在函数$f(D)$上加入拉普拉斯噪声,得到输出函数A为

$$A(D)=f(D)+\mathrm{Lap}\left(\frac{GS(f)}{\varepsilon}\right)^k \tag{2}$$

拉普拉斯分布的均值为0,并且有密度函数$p(x,b)=\frac{1}{2b}\exp\left(-\frac{|x|}{b}\right)$。

(2) 敏感度

在拉普拉斯机制中,敏感度是影响隐私保护强度的一个重要因素,其定义如下。

定义2[10] 若给定查询函数f,$f(D)\in$Range(f),$f(D)$表示查询函数f在数据集D上的查询结果,Range(f)表示f的查询结果取值范围,则函数的敏感度值为

$$S(F)=\max(|f(D)-f(D')|) \tag{3}$$

式中D,D'为类似于差分隐私定义中的相邻数据集。

(3) 隐私预算与敏感度

差分隐私预算ε与敏感度的关系定理如下。

定理1[12] 给定函数集F,其敏感度为$S(F)$,K表示向F的输出结果添加独立噪声的算法,若添加的噪声满足取值为$S(F)/\varepsilon$的拉普拉斯分布,则算法K满足ε-差分隐私。

1.2 生成对抗网络

GAN[13]是一种无监督的深度学习算法,其体系结构通常包括两个模块:生成器G和判别器D。G

的目标是通过学习从一个潜在分布p_z到真实数据分布p_{data}的映射，生成尽可能看似来自p_{data}的假样本来“欺骗”判别器D；而D的目标是区分虚假样本和真实样本之间的区别。由于G与D两者的目标相反，故GAN的训练可以看作是一个博弈过程。在博弈过程中，算法分别依次优化D与G，直到两者收敛。因此，整体优化目标函数可写为

$$\min_{\theta}\max_{w} E_{x\sim p_{\text{data}}}[\log D_w(x)]+E_{z\sim p_z}[\log(1-D_w(G_\theta(z)))] \tag{4}$$

尽管GAN结构简单，但是最初版本的GAN既不稳定，训练效率也不高。在GAN的后续研究工作中，新的训练机制和网络模型被提出，以不断地提高训练的稳定性和收敛速度。例如，Martin等[14]提出的沃塞斯坦对抗性网络（Wasserstein GAN, WGAN）和改进的WGAN训练机制解决了原始GAN存在的训练效率低下与模式坍缩的问题。不同于原始GAN的目标函数最小化真实分布与生成分布之间的JS-散度（Jensen-Shannon），WGAN通过Wasserstein距离来衡量两个分布之间的差异，并采用了一种梯度惩罚技术来限制判决器的Lipschitz条件，使目标函数变为

$$\arg\min_{\theta} D_w(G_\theta(z))-D_w(x)+\lambda(\|\nabla_{\hat{x}}D_w(\hat{x})\|_2-1)^2 \tag{5}$$

式中：$\hat{x}=\alpha x+(1-\alpha)G_\theta(z)$，$\alpha$采样于$[0,1]$。正则项将$D$的梯度范数逼近于1。该机制可以使得网络的训练更加稳定，效率更高。

若无特别声明，本文中使用的生成对抗网络模型均指改进的WGAN。

2 模型的设计与实现

本文设计的支持图像差分隐私保护的生成对抗网络模型（IDP-GAN）的重点在于生成对抗网络结构中噪声添加位置的选取和隐私预算的分配，下面阐述模型的设计思路和实现过程。

2.1 拉普拉斯噪声添加位置

本文选取GAN结构中的判别器D作为噪声机制添加的位置。因为在生成对抗网络中，只有判别器D能够接触到真实的数据，对判别器D扰动足以控制GAN生成数据的隐私。此外，不同于生成器G的复杂网络结构，判别器D仅需要较简单的网络结构和较少的参数，这样更容易计算隐私损失。更重要的是，根据差分隐私的后处理性，对判别器D实现差分隐私，生成器G的参数也同样满足差分隐私保护，且对差分隐私输出的任何计算不会增加隐私损失（计算是指对生成器参数的计算，输出指判别器输出的差分隐私保护参数）。

2.2 差分隐私生成对抗网络的构建

IDP-GAN的设计核心在于构建差分隐私生成对抗网络，其关键步骤包括差分隐私神经网络的构建以及差分隐私损失函数的构造，均以单次引入噪声实现差分隐私，减少隐私损失。

2.2.1 差分隐私神经网络的构建

本文通过将拉普拉斯噪声加入输入特征，以构建差分隐私神经网络。这里的输入特征指的是生成对抗网络判别器中神经网络的仿射变换层的特征。

首先，生成对抗网络的判别器中神经网络的仿射变换层可表示为

$$h_{x_i}(\boldsymbol{W})=\boldsymbol{x}_i\boldsymbol{W}^{\mathrm{T}}+b \tag{6}$$

式中：b为静态偏差，$\boldsymbol{W}$为h的参数。在经过L批次训练后，仿射变换可表示为

$$h_L(\boldsymbol{W})=\sum_{x_i\in L}(\boldsymbol{x}_i\boldsymbol{W}^{\mathrm{T}}+b) \tag{7}$$

得到单个神经元$h_L(\boldsymbol{W})$的表示后，通过将隐私预算为ε_1拉普拉斯噪声注入每个神经元$h_L(\boldsymbol{W})\in h_0$

的和输入特征$\boldsymbol{x}_i$和偏差b,来扰动仿射变换层h_0。拉普拉斯噪声分布为$\frac{1}{|L|}\mathrm{Lap}(\Delta h_0/\varepsilon_1)$,其中$\Delta h_0=2\sum_{h\in h_0}d$。扰动后的输入特征和偏差分别表示为$\overline{\boldsymbol{x}}_i$和$\overline{b}$。给定随机训练批次$L$,扰动后的差分隐私仿射变换层记为$\overline{h}_{0L}$,其中每个隐藏神经元$\overline{h}_L(\boldsymbol{W})$表示为

$$\overline{h}_{0L}(\boldsymbol{W}_0)=\left\{\overline{h}_L(\boldsymbol{W})\right\}_{h\in h_0}$$
$$\text{s.t. }\overline{h}_L(\boldsymbol{W})=\sum_{\boldsymbol{x}_i\in L}(\overline{\boldsymbol{x}}_i\boldsymbol{W}^{\mathrm{T}}+\overline{b}) \tag{8}$$

然后,将隐藏层$\{h_1,\cdots,h_k\}$堆叠在差分隐私仿射变换层$\overline{h}_{0L}$以构成差分隐私深度神经网络,如图1所示。

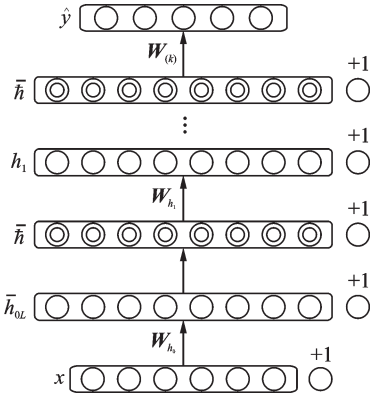

图1 差分隐私神经网络的示例

Fig.1 An instance of differentially private neural networks

$\overline{h}_{0L}$表示扰动后的仿射变换层,从h_1至h_k层的计算均是基于差分隐私仿射变换层$\overline{h}_{0L}$完成的,没有接触到原始数据中的任何信息。因此,计算过程不会公开原始数据集中的任何隐私信息。在每次堆叠操作之前,通过归一化层$\overline{h}$配置非线性函数进行激活。

2.2.2 差分隐私损失函数的构造

差分隐私神经网络的输出层产生预测标签Y。由于预测标签y_i与原始数据集具有相同的隐私性,因此在输出层也需要实现差分隐私以保护标签y_i。首先,基于泰勒展开(Taylor expansion)[15]对损失函数的多项式近似进行推导。然后,将拉普拉斯噪声注入损失函数$F_L(\theta)$的系数中,以在每个训练批次L上实现差分隐私。

差分隐私神经网络模型的输出变量$\{\hat{y}_1,\cdots,\hat{y}_M\}$通过加权$\boldsymbol{W}(k)$与顶层的归一化层进行全连接,在激活函数作用下,给定第l个输出变量$\hat{y}_l$和输入x_i,则输出标签可表示为

$$\hat{y}_{il}=\sigma(\overline{h}_{\boldsymbol{x}_i(k)}\boldsymbol{W}_{l(k)}^{\mathrm{T}}) \tag{9}$$

式中$\overline{h}_{\boldsymbol{x}_i(k)}$表示神经网络中隐藏神经元的状态。

本文以应用最广的交叉熵损失函数[16]为例进行多项式推导,也可以选取均方误差等其他损失函数。整个神经网络模型的交叉熵损失函数的计算过程如下。

$$F_L(\theta)=-\sum_{l=1}^{M}\sum_{\boldsymbol{x}_i\in L}(y_{il}\log\hat{y}_{il}+(1-y_{il})\log(1-\hat{y}_{il}))=$$
$$-\sum_{l=1}^{M}\sum_{\boldsymbol{x}_i\in L}(y_{il}\log(1+\mathrm{e}^{-\overline{h}}\boldsymbol{x}_{i(k)}\boldsymbol{W}_{l(k)}^{\mathrm{T}})+(1-y_{il})\log(1+\mathrm{e}^{\overline{b}}\boldsymbol{x}_{i(k)}\boldsymbol{W}_{l(k)}^{\mathrm{T}})) \tag{10}$$

基于泰勒展开,$F_L(\theta)$的多项式近似可表示为

$$\hat{F}_L=\sum_{l=1}^{M}\sum_{\boldsymbol{x}_i\in L}\sum_{q=1}^{2}\sum_{R=0}^{2}\frac{f_{ql}^{(R)}(0)}{R!}(\overline{h}_{\boldsymbol{x}_i(k)}\boldsymbol{W}_{l(k)}^{\mathrm{T}})^R=$$
$$-\sum_{l=1}^{M}\sum_{\boldsymbol{x}_i\in L}(y_{il}\log(1+\mathrm{e}^{-\overline{h}}\boldsymbol{x}_{i(k)}\boldsymbol{W}_{l(k)}^{\mathrm{T}})+(1-y_{il})\log(1+\mathrm{e}^{\overline{b}}\boldsymbol{x}_{i(k)}\boldsymbol{W}_{l(k)}^{\mathrm{T}})) \tag{11}$$

式中:$\forall l\in[1,M]$:$f_{1l}(z)=y_{il}\log(1+\mathrm{e}^{-z})$,且$f_{2l}(z)=(1-y_{il})\log(1+\mathrm{e}^{z})$。

为了实现差分隐私保护,本文对泰勒展式的每项系数插入拉普拉斯噪声,从而达到扰动损失函数的目的。因此,将y_{il}对应的泰勒展式系数定义为$\left\{\phi_{l\boldsymbol{x}_i}^R\right\}$,$R\in D$。给定隐私预算为$\varepsilon_2$,则加入的拉普拉斯噪声为$\frac{1}{L}\mathrm{Lap}(\Delta F/\varepsilon_2)$,$\Delta F=M(|\overline{h}_{(k)}|+\frac{1}{4}|\overline{h}_{(k)}|^2)$。然后,将扰动后的系数和损失函数分别表示为$\left\{\overline{\phi}_{l\boldsymbol{x}_i}^R\right\}$

和$\overline{F}_L(\theta_t)$,这样扰动后的损失函数就无法直接接触原始的预测标签。

2.2.3 隐私损失的计算

在IDP-GAN的差分隐私实现机制中,在访问原始数据的每个计算任务中都会执行差分隐私保护。但是,拉普拉斯噪声仅需要被单次注入到IDP-GAN模型中,作为预处理步骤在放射变换层$\overline{h}_{0L}$和损失函数$\overline{F}_L(\theta)$中实现差分隐私。此后,训练阶段将不再访问原始数据,隐私预算消耗不会在每个训练步骤中累积。因此,隐私预算消耗与训练时期的数量无关,总隐私预算$\varepsilon=\varepsilon_1+\varepsilon_2$。

2.3 模型的实现

根据上文的支持图数据差分隐私保护的生成对抗网络模型(IDP-GAN)设计思想及流程分析,本节给出IDP-GAN模型的伪代码(如算法1所示)。在介绍具体实现过程之前,先介绍在IDP-GAN模型构造中用到的符号表示。模型的损失函数记为$F(\theta)$,这里的θ表示参数。模型随机训练T批次,每次训练批次L,批次L是D中具有预定批次大小$|L|$的训练样本的随机集合。最后使用Adam算法优化损失函数。

算法1 IDP-GAN

输入:图数据集D;隐藏层H;损失函数$F(\theta)$;隐私预算ε_1和ε_2;Adam超参数α, β_1和β_2;判别器的训练迭代次数T;随机训练批次大小$|L|$;学习率η_t;随机噪声分布p_z。

输出:差分隐私生成器G。

(1) 初始化θ_0

(2) For $t\in|T|$, $l\in|L|$ do,$x_i\sim D$, $z\sim p_z$

(3) $\overline{\boldsymbol{x}}_i\leftarrow\boldsymbol{x}_i+\frac{1}{|L|}\mathrm{Lap}(\frac{\Delta h_0}{\varepsilon_1})$

(4) $\overline{b}\leftarrow b+\frac{1}{|L|}\mathrm{Lap}(\frac{\Delta h_0}{\varepsilon_1})$ //扰动判别器的仿射变换层

(5) $\overline{h}_{0L}(\boldsymbol{W}_0)=\{\overline{h}_L(\boldsymbol{W})\}_{h\in h_0}$

(6) $\overline{h}_L(\boldsymbol{W})=\sum_{x_i\in L}(\overline{\boldsymbol{x}}_i\boldsymbol{W}^{\mathrm{T}}+\overline{b})$ //构建差分隐私神经网络

(7) $\overline{\Phi}_{lx_i}^{(R)}\leftarrow\Phi_{lx_i}^{(R)}+\frac{1}{|L|}\mathrm{Lap}(\frac{\Delta_F}{\varepsilon_2})$

(8) $\overline{F}_L=\sum_{l=1}^{M}\sum_{x_i\in L}\sum_{R=0}^{2}\overline{\Phi}_{lx_i}^{(R)}(\overline{h}_{x_i(k)}\boldsymbol{W}_{l(k)}^{\mathrm{T}})^R$ //构建差分隐私损失函数

(9) $\theta_{t+1}\leftarrow\theta_t-\eta_t\frac{1}{|L|}\nabla_{\theta_t}\overline{F}_L(\theta)$ //计算判别器的梯度下降

(10) $w\leftarrow\mathrm{Adam}(\frac{1}{L}\sum_{l=1}^{L}\theta_{t+1},w,\alpha,\beta_1,\beta_2)$ //更新判别器

(11) $\theta\leftarrow\mathrm{Adam}(\nabla_\theta\frac{1}{L}\sum_{l=1}^{L}-D(G(z)),\theta,\alpha,\beta_1,\beta_2)$ //更新生成器

(12) $\varepsilon\leftarrow\varepsilon_1+\varepsilon_2$ //计算隐私损失

(13) Return差分隐私生成器G

3 实验结果及分析

通过实验对比及统计的实验数据,从不同角度分析了支持图数据差分隐私保护的生成对抗网络模

型(IDP-GAN)的有效性。

3.1 实验环境及数据集介绍

本文在两个标准数据集(MNIST数据集[17]和CelebA数据集[18])上进行了大量的实验,通过评估IDP-GAN生成数据的质量和隐私级别,验证的IDP-GAN性能。

MNIST数据集由大小为28像素×28像素的手写数字图像组成,分为60 KB和10 KB测试样本;CelebA数据集包含20万名人的脸部图像大小为48像素×48像素,每个图像有40个属性注释。

本文分别将MNIST和CelebA数据集的训练数据(如果不考虑标签信息,则训练数据为整个数据集)按照2:98的比例分为公有数据集D_{pub}和隐私数据集D_{pri}。

实验环境为Ubuntu 16.04.5 LTS操作系统,GTX 1080 Ti和NVIDIA TESLA K40C两块显卡,实验利用TensorFlow版本1.9.0和Keras版本2.2.0的框架,以及Python版本3.6.5实现。

本文使用Leaky ReLUs作为鉴别器上的激活函数,并使用ReLU作为生成器上的激活函数,因此激活函数导数的界限$B_\sigma' \leqslant 1$。在生成器和判别器的权重上加入L_2-regularization,以在学习过程中降低模型复杂度和不稳定程度,IDP-GAN超参数设置如表1所示。

表1 IDP-GAN模型的超参数设置

Table 1 Hyperparameter of IDP-GAN

超参数	说明	取值
α_d	判别器学习率	5.0×10^{-5}
α_g	生成器学习率	5.0×10^{-5}
δ	噪声级别	10^{-5}
B_σ'	激活函数导数的界限	$\leqslant 1$

3.2 性能指标

3.2.1 Inception score

Salimans等[19]提出了用Inception score来衡量GAN生成数据的质量。理论上,生成器G的Inception score定义为

$$S(G)=\exp(E_{x\sim G(z)}KL(P_r(y|x)\|P_r(y))) \tag{12}$$

式中:x为由G生成的样本,而$P_r(y|x)$为预先训练的分类器的条件分布,用于预测x的标签y。如果x与真实样本相似,则期望$P_r(y|x)$的熵很小。$P_r(y)$是y的边缘分布。如果G能够生成一组不同的样本,则期望$P_r(y)$的熵很大。因此,通过测量两个分布的KL-散度(KL-divergence),$S(G)$可以评估生产数据的质量和多样性。对于MNIST和CelebA数据集,使用完整的训练集来训练基线分类器以估算$P_r(y|x)$,调整分类器以在验证集上评估性能。简而言之,Inception score的值越大,表示生成模型生成图像的质量和多样性就越高。

3.2.2 推理攻击的精度

为评估IDP-GAN的隐私级别,本文通过成员推理攻击[20]来衡量将生成图像用于训练模型会造成的成员泄露风险。当攻击者获得某项记录并且该记录已用于训练特定模型,则表明该模型存在信息泄漏的风险,且推理攻击的精度越低,隐私的级别越高。

3.3 相关实验及结果分析

在本文的实验中,以在随机梯度下降的计算中添加相同噪声的dp-GAN[9]作为对比模型进行IDP-GAN模型的性能测试。

3.3.1 隐私级别与生成图像质量的关系

为验证隐私级别与IDP-GAN的输出图像质量之间的关系,在MNIST和CelebA数据集上进行了大量的实验。在这些实验中,根据dp-GAN中隐私预算的选择,在实验中选择一些相对较大隐私预算的(范围从0.3到11)来评估IDP-GAN的性能,对应于3个不同的ε值,IDP-GAN生成的图像分别如图2和图3所示。

(a) $\varepsilon=2, \delta=10^{-5}$ (b) $\varepsilon=4, \delta=10^{-5}$ (c) $\varepsilon=8, \delta=10^{-5}$

图2 基于MNIST的3种不同ε的生成图像样例

Fig.2 Synthetic samples with three different ε on MNIST dataset

(a) $\varepsilon=2, \delta=10^{-5}$ (b) $\varepsilon=4, \delta=10^{-5}$ (c) $\varepsilon=8, \delta=10^{-5}$

图3 基于CelebA的3种不同ε的生成图像样例

Fig.3 Synthetic samples with three different ε on CelebA dataset

实验结果表明,IDP-GAN可以生成视觉上清晰的图像,图像的失真是由注入的噪声造成的而不是因为训练图像的质量问题。将生成的图像与其相邻图像进行对比,可以明显地看到,IDP-GAN不仅是为了学习训练样本,而且还能够生成具有独特细节的图像。更重要的是,图2和图3中生成的图像显示,当所有其他条件都相同时,噪声的方差越大,生成图像的模糊性就越大。在IDP-GAN中,任何获得合成图像的使用者几乎都无法知道训练过程中是否涉及某个数据,在这种情况下无法重建训练图像,因此可以保护原始数据的隐私。尽管噪声减小会提高生成图像的质量,但模型的安全性通常也会减小。因此,选择一个合理的ε(添加更少的噪声),能够避免对生成的数据造成太大影响。这也表明IDP-GAN成功解决了前文提到的隐私问题,并且在较大范围内调整隐私级别ε可以确保生成图像的高质量。

3.3.2 生成图像的质量评估

下面通过实验,根据Inception score将生成数据与真实数据进行比较,对IDP-GAN生成图像的质量进行评估,使用完整的训练集来训练基线分类器以估算$P_r(y\mid x)$,调整分类器以在验证集上评估性能。Inception score的值越大,表示生成模型具有生成更高质量且具备多样性的图像的能力。

为了评估IDP-GAN生成图像的质量,通过在MNIST和CelebA数据集上进行实验并与dp-GAN和无隐私保护的WGAN进行对比。图4比较了具有不同隐私预算条件下3种生成模型的生成图像和真实图像的Inception score。根据Inception score的统计特性,从图4可以看出,IDP-GAN生成图像的质量优于dp-GAN。

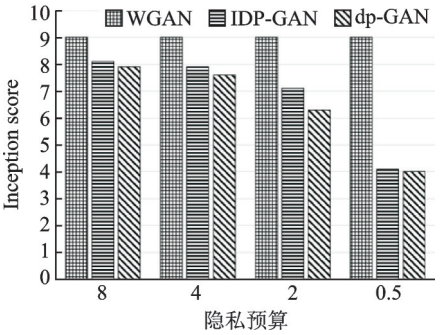

图4 3种生成模型的Inception score对比

Fig.4 Inception scores comparison of three generative models

3.3.3 隐私级别的评估

为评估IDP-GAN的隐私级别,通过成员推理攻击来衡量将生成图像用于训练模型会造成的成员泄露风险。当攻击者获得某项记录并且该记录已用于训练特定模型时,则表明该模型存在隐私泄漏的风险。在攻击实验中,使用CelebA数据集验证隐私预算对攻击准确性的影响。图5显示了在IDP-GAN中使用不同隐私预算和不同规模的数据集训练时,实现成员推理攻击的精度。

从图5可以看出,数据集的大小和隐私预算的大小均与攻击精度成正相关。隐私预算越小,攻击精度就越低。在差分隐私理论中,隐私预算反映了隐私保护的程度,值越小说明隐私保护程度越高,当隐私预算设2时,即使有大规模的数据集,攻击精度最高也仅为0.52。因此在隐私预算较小时,攻击者无法准确地推断出目标模型的分布,从而证明IDP-GAN可以缓解生成模型的信息泄漏。

为进一步评估IDP-GAN的隐私性,本文将IDP-GAN与dp-GAN和WGAN(无隐私保护)这两种现有解决方案下CelebA数据集实现成员推理攻击的精度进行比较。图6显示了在不同解决方案下CelebA数据集实现成员推理攻击的精度。从图6可以观察到,IDP-GAN抵御成员推理攻击的能力优于dp-GAN和WGAN。

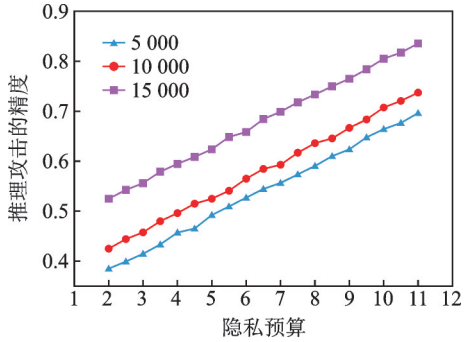

图5 基于不同大小的CelebA数据集在不同的隐私预算下实现推理攻击的精度

Fig.5 Precision of the inference attack for CelebA dataset with different sizes of datasets

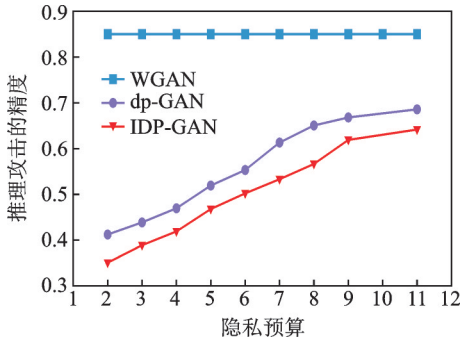

图6 不同解决方案下CelebA数据集实现推理攻击的精度

Fig.6 Precision of the inference attack for CelebA dataset under different solutions

4 结束语

本文立足于研究数据挖掘过程中隐私保护强度与数据可用性之间的平衡问题,针对现有的隐私保护和数据挖掘结合的方法展开深入研究,提出了一种支持图像差分隐私保护的生成对抗网络模型(IDP-GAN)。IDP-GAN将拉普拉斯噪声合理地分配到判别器以提高生成对抗网络的隐私性。更重要的是,不同于在“梯度”中引入相同噪声量的隐私保护机制会造成隐私预算冗余,IDP-GAN仅在判别器的仿射变换层和损失函数的多项式近似系数中单次加入拉普拉斯噪声,减少了隐私预算的消耗,大幅提升了差分隐私生成对抗网络生成图像的精度。

如何在数据挖掘模型训练过程中保护隐私信息的研究将具有广阔的研究前景和实用价值。本文聚焦于图像数据集实现了支持图数据差分隐私保护的生成对抗网络,然而差分隐私与其他类型的语义丰富数据(如文本、音频等模型)生成模型的结合也是值得深入研究的。

参考文献:

[1] 冯登国. 大数据安全与隐私保护[J]. 计算机学报, 2014, 37(1): 246-258.

FENG Dengguo. Big data security and privacy protection[J]. Chinese Journal of Computers, 2014, 37(1): 246-258.

[2] 方滨兴, 贾焰, 李爱平, 等. 大数据隐私保护技术综述[J]. 大数据, 2016, 2(1): 1-18.
FANG Binxing, JIA Yan, LI Aiping, et al. Privacy preservation in big data:A survey[J]. Big Data, 2016, 2(1): 1-18.

[3] ELISA Bertino. Big data security and privacy[C]// Proceedings of 2016 IEEE International Conference on Big Data. [S.l.]: IEEE, 2016.

[4] GOODFELLOW I J. On distinguishability criteria for estimating generative models[J]. Statistics, 2014.

[5] HITAJ B, ATENIESE G, PEREZ-CRUZ F. Deep models under the GAN: Information leakage from collaborative deep learning[C]//Proceedings of the ACM SIGSAC Conference. [S.l.]: ACM, 2017.

[6] JAIN P, GYANCHANDANI M , KHARE N . Differential privacy: Its technological prescriptive using big data[J]. Journal of Big Data, 2018, 5(1): 15.

[7] PAPERNOT N, MCDANIEL P, WU X, et al. Distillation as a defense to adversarial perturbations against deep neural networks[C]//Proceedings of 2016 IEEE Symposium on Security and Privacy (SP). [S.l.]: IEEE, 2016.

[8] ABADI M, CHU A, GOODFELLOW I, et al. Deep learning with differential privacy[C]//Proceedings of the 2016 ACM SIGSAC Conference. [S.l.]: ACM, 2016.

[9] ACS G, MELIS L, CASTELLUCCIA C, et al. Differentially private mixture of generative neural networks[J]. IEEE Transactions on Knowledge and Data Engineering, 2019,31(6): 1109-1121.

[10] DWORK C. Differential privacy in new settings[C]//Proceedings of the 21st Annual ACM-SIAM Symposium on Discrete Algorithms. Austin, Texas, USA: ACM, 2010.

[11] DWORK C. Differential privacy[J]. Lecture Notes in Computer Science, 2006, 26(2): 1-12.

[12] VISWANATH P. The optimal noise-adding mechanism in differential privacy[J]. IEEE Transactions on Information Theory, 2016, 62(2): 925-951.

[13] MAKHZANI A , SHLENS J , JAITLY N , et al. Adversarial autoencoders[C]//Proceedings of the International Conference on Learning Representations. San Juan, USA: [s.n.], 2016.

[14] CUI S, JIANG Y. Effective Lipschitz constraint enforcement for Wasserstein GAN training[C]//Proceedings of 2017 the 2nd IEEE International Conference on Computational Intelligence and Applications(ICCIA). Beijing, China: IEEE, 2017: 2-17.

[15] KANWAL R P, LIU K C. A Taylor expansion approach for solving integral equations[J]. International Journal of Mathematical Education in Science & Technology, 1989, 20(3): 411-414.

[16] GOLOVKO V A. Deep learning: An overview and main paradigms[J]. Optical Memory and Neural Networks, 2017, 26(1): 1-17.

[17] DENG L. The MNIST database of handwritten digit images for machine learning research [Best of the web][J]. Signal Processing Magazine, IEEE, 2012, 29(6): 141-142.

[18] KAZEMI H, IRANMANESH M, DABOUEI A, et al. Facial attributes guided deep sketch-to-photo synthesis[C]// Proceedings of IEEE Winter Applications of Computer Vision Workshops. [S.l.]: IEEE, 2018.

[19] RIKIYA Y, MIZUHO N, GIAN D R K, et al. Convolutional neural networks: An overview and application in radiology[J]. Insights into Imaging, 2018, 9: 611-629.

[20] SONG L, SHOKRI R, MITTAL P. Membership inference attacks against adversarially robust deep learning models[C]// Proceedings of 2019 IEEE Security and Privacy Workshops(SPW). [S.l.]: IEEE, 2019.

作者简介:

杨云鹿(1994-),女,硕士研究生,研究方向:大数据隐私保护、大数据智能化处理、深度学习等,E-mail: yangyunlu@caict.ac.cn。

周亚建(1971-),**通信作者**,男,博士,副教授,研究方向:文本分类、密文检索、大数据分析、数据隐私保护等。

宁华(1975-),女,博士,高级工程师,研究方向:移动互联网安全、个人信息保护等。

(编辑:王静)