基于 DCGAN 反馈的深度差分隐私保护方法

毛典辉[1,2]，李子沁[1,2]，蔡 强[1,2]，薛子育[1,2]
(1. 北京工商大学计算机与信息工程学院，食品安全大数据技术北京市重点实验，北京 100048；
2. 北京工商大学农产品质量安全追溯技术及应用国家工程实验室，北京 100048)

摘 要： 为了防止攻击者在深度学习模型应用过程中利用生成式对抗网络(generative adversarial networks，GAN)等技术还原出训练集中的数据，保护训练数据集中用户的敏感信息，提出一个基于深度卷积生成式对抗网络(deep convolutional generative adversarial networks，DCGAN)反馈的深度差分隐私保护方法. 该方法在深度网络参数优化计算时结合差分隐私理论添加噪声数据，基于差分隐私与高斯分布可组合特点，计算深度网络每一层的隐私预算，在随机梯度下降(stochastic gradient descent，SGD)计算中添加高斯噪声使之总体隐私预算最小；利用 DCGAN 生成数据选取可能得到的最优结果，通过对比攻击结果和原始数据之间的差别调节深度差分隐私模型参数，实现训练数据集可用性与隐私保护度的平衡. 实验结果表明，该方法针对训练数据集中的敏感信息具有较高的隐私保护能力.

关键词： 训练数据集保护；差分隐私；深度学习；图像生成；深度卷积生成式对抗网络(DCGAN)
中图分类号： U 461；TP 308 **文献标志码：** A **文章编号：** 0254－0037(2018)06－0870－08
doi： 10.11936/bjutxb2017070017

Tickling Deep Differential Privacy Protection Method Based on DCGAN

MAO Dianhui[1,2], LI Ziqin[1,2], CAI Qiang[1,2], XUE Ziyu[1,2]
(1. Beijing Key Laboratory of Big Data Technology for Food Safety, School of Computer and Information Engineering, Beijing Technology and Business University, Beijing 100048, China;
2. National Engineering Laboratory for Quality and Safety Traceability Technology and Application of Agricultural Products, Beijing Technology and Business University, Beijing 100048, China)

Abstract: To prevent attackers from using generative adversarial networks (GAN) and other technologies in the application of deep learning model to restore the data in the training dataset, and to protect the sensitive information of users in the training dataset, a ticking deep differential privacy protection method was proposed based on DCGAN. The noise data was added the differential privacy theory when optimizing the in-depth network parameters in this method. Then the privacy budget of each layer of the deep network was calculated in a stochastic gradient descent (SGD), which based on the combination of differential privacy and Gaussian distribution, Gaussian noise was added to minimize the total privacy budget in the stochastic gradient descent calculation. And then the optimal result that the attacker may obtain was generated by using DCGAN. Finally, in order to achieve balance between data availability and privacy protection, the difference among the attack result and the original data was used to adjust the deep differential privacy model. The results show that this method has high privacy protection ability for

收稿日期： 2017-07-12
基金项目： 北京市科技计划资助项目(Z161100001616004)；北京工商大学两科基金培育项目(LKJJ2017-13)；教育部人文社会科学基金资助项目(17YJCZH127)
作者简介： 毛典辉(1979—)，男，副教授，主要从事数据发布与信息隐藏、时空数据挖掘方面的研究，E-mail: maodh@th.btbu.edu.cn

sensitive information in training dataset.

Key words: training dataset protection; differential privacy; deep learning; image generation; deep convolutional generative adversarial networks (DCGAN)

近年来，深度学习在目标检测和计算机视觉、自然语言处理、语音识别和语义分析等领域成效卓然，受到了越来越多研究者的关注[1]. 深度学习通过神经网络的分层处理，将低层特征组合成更加抽象的高层表示属性类别或特征，以发现数据的分布式特征表示[2]，其模型性能与训练数据集的规模和质量密切相关，而训练数据集中通常包含较多的敏感信息，攻击者通过一定的攻击手段可以还原出训练数据集，从而使得用户隐私信息泄露. 例如：公安机关发布犯罪嫌疑人识别模型，其训练数据集包含全国人口图像信息，当攻击者使用某一攻击手段还原出训练数据集中的图像时，会使个人敏感信息泄露. 因此，如何在不泄露个人敏感信息的前提下提升数据可用性，是当前深度学习应用面临的主要问题，将极大影响深度学习未来的发展.

目前，关于敏感信息保护问题的研究的前提主要基于攻击者对用户背景知识的掌握程度，在此条件下攻击者可以进行身份链接攻击、属性链接攻击、成员链接攻击等隐私攻击，因此，相关学者提出了 K-匿名[3]、L-多样性[4]以及相关的方法[5]. 该类方法通过泛化或抑制用户敏感属性并修改数据集中的原始信息的策略，从而达到保护用户隐私的目的，而深度学习模型主要通过提取并抽象训练数据集中的特征，并不改变数据集的原始信息，因此，与传统方法融合时存在较大难度. 2015 年，Reza 等[6]设计了一个深层神经网络分布式系统实现了训练数据集的隐私保护，该系统允许参与者在自己的数据集上独立训练，并在训练期间选择性地共享模型关键参数子集，该过程使得参与者可以保留其各自数据的隐私，同时仍然受益于其他参与者的模型，从而提高他们的学习准确性. 但 Matt 等[7]模型反演攻击可以利用去噪自编码网络还原训练数据集中的原始信息；Ian 等[8]利用生成式对抗网络(generative adversarial networks，GAN)生成与训练数据集相近的数据. 为了解决模型反演攻击，Nhat 等[9]提出差分隐私自编码(ε-differential private autoencoder)方法，该方法利用差分隐私理论来扰乱深度自动编码器中的目标函数，在数据重建过程中添加噪声从而使得训练数据集得以保护. Nicolas 等[10]提出了一种教师-学生模式的深度网络隐私保护方法，该方法包含多个由敏感信息数据集训练的教师深度模型以及一个由 GAN 模型生成的用于预测的学生模型，学生模型是利用教师模型在投票时结合差分隐私理论选出较优的预测结果，使用者利用学生模型进行预测，教师模型不公布，从而达到保护训练数据集的目的. 上述过程均将深度学习模型训练过程视作黑盒子，仅从模型外部添加隐私保护机制，其保护效果可控性略显不足.

针对上述深度学习隐私保护过程的不透明性，本文在无需考虑攻击者所拥有的任何可能的背景知识的情况下，在深度学习网络训练过程中引入差分隐私理论，设计了一个基于深度卷积生成式对抗网络(deep convolutional generative adversarial networks，DCGAN)反馈的深度差分隐私保护方法.

1　相关定义

差分隐私技术是 2006 年微软研究院提出的一种基于数据失真的隐私保护方法[11]，该方法建立在坚实的数学基础之上，对隐私保护进行了严格的定义并提供了量化评估方法，使得不同参数处理下的数据集所提供的隐私保护水平具有可比较性[12]. 其基本思想是通过对原始数据、原始数据的转换或者统计结果添加噪声来达到隐私保护效果. 该保护方法可以确保在某一数据集中插入或者删除一条记录的操作不会影响任何计算的输出结果. 另外，该保护模型不关心攻击者所具有的背景知识，即使攻击者已经掌握除某一条记录之外的所有记录的信息.

定义 1　(ε,δ)-差分隐私　给定 2 个训练数据集 D、D'，其中 D 与 D' 二者之间最多相差一条记录，给定一个算法 A，A 的取值范围为 Range(A)，若算法 A 在数据集 D 和 D' 上任意输出结果 R($R \in$ Range(A)) 满足 $\Pr[A(D') = R] \leqslant e^{\varepsilon} \times \Pr[A(D) = R] + \delta$，其中，$\varepsilon$ 为隐私保护程度，δ 为误差值，且均为非负值，则算法 A 满足(ε,δ)-差分隐私，而 ε 值越小说明算法 A 的隐私保护程度越好[13].

差分隐私理论与深度学习模型结合的方式主要有 2 种：一种是将深度学习模型视为黑盒子，在深度学习模型训练好的最终参数中添加噪声数据；另一种是将深度学习模型视为白盒子，在模型训练过程

中的参数优化阶段添加噪声数据. 前者因某些深度学习模型特征参数依赖训练数据,对参数添加过多噪声将破坏模型的可用性,当噪声数据添加过少,则无法实现训练数据集的隐私保护目的. 因此,本文采用在深度学习模型训练过程中引入差分隐私理论,提出一种(ε,δ)-深度差分隐私保护模型,其形式化定义如下.

定义 2 (ε,δ)-深度差分隐私 针对一个深度学习网络,给定一个算法 A,在每次优化计算过程中添加分布函数 f 的高斯噪声,即 $A(d) \triangleq f(d) + \mathcal{N}(0, S_f^2\sigma^2)$. 给定 2 个数据集 D、D',其中 D 与 D' 二者之间至多相差一条记录,若 f 满足 $\delta \geqslant \frac{4}{5} \cdot \exp(-(\sigma\varepsilon)^2/2)$ 且 $\varepsilon < 1$[14],则算法 A 满足(ε,δ)-深度差分隐私. 其中 $\mathcal{N}(0, S_f^2\sigma^2)$ 是均值为 0,方差为 $S_f\sigma$,敏感系数为 S_f 的高斯分布.

2 算法思路与实现

上述(ε,δ)-深度差分隐私保护模型在实现过程中存在 2 个难点:1) 深度学习模型结构中噪声机制添加位置的选取;2) 隐私预算的合理分配. 为了解决该问题,本文的思路为:首先,在深度学习网络训练中的参数优化阶段,基于差分隐私与高斯分布可组合的特点,计算深度网络每一层的隐私预算,在随机梯度下降计算中添加高斯噪声使之总体隐私预算最小;然后,以深度差分隐私算法处理的数据为基础,提取训练后特征,利用 DCGAN 生成数据,并从中选出最接近真实数据集的攻击结果;最后,比较攻击结果和原始数据集的相似度,若相似度超过设定阈值,则重新对深度差分隐私模型进行参数调优,直到满足条件为止. 其算法实现过程如图 1 所示,该算法可以实现在深度学习训练过程中根据用户意愿设置的参数实现保护用户敏感信息.

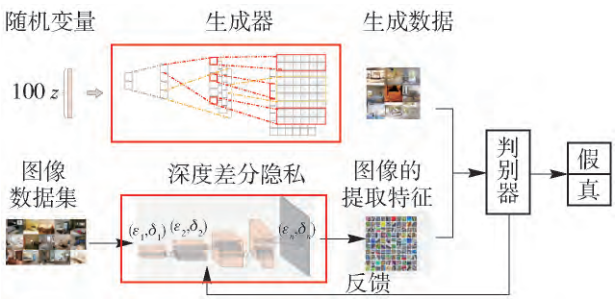


图 1 基于 DCGAN 反馈的深度差分隐私保护框架

Fig. 1 Tickling deep differential privacy protection method based on DCGAN

2.1 深度差分隐私算法实现

深度学习模型的基本结构一般是由输入层、多层隐含层及输出层构成的神经网络模型,其训练过程通过逐层特征变换将样本在原空间的特征表示转换到新特征空间,从而使分类和预测更加容易[15]. 在与差分隐私理论融合的过程中,一个复杂的敏感信息保护问题通常需要多次应用差分隐私保护算法才能得以解决,在这种情况下,为了保证使整个过程的隐私保护水平控制在给定的预算 ε 之内,需要合理地将全部预算分配到整个算法的各个步骤中.

性质 1 设有算法 $M_1, M_2, \cdots, M_n$,其隐私保护预算分别为 $\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_n$,那么对于同一数据集 D,由这些算法构成的组合算法 $M(M_1(D), M_2(D), \cdots, M_n(D))$ 提供($\sum_{i=1}^{n} \varepsilon_i$)-差分隐私保护[16].

该性质表明,一个差分隐私保护算法序列构成的组合算法,其提供的隐私保护水平为全部预算的总和,该性质也称为序列组合性.

性质 2 设有算法 $M_1, M_2, \cdots, M_n$,其隐私保护预算分别为 $\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_n$,那么对于不相交的数据集 $D_1, D_2, \cdots, D_n$,由这些算法构成的组合算法 $M(M_1(D), M_2(D), \cdots, M_n(D))$ 提供($\max \varepsilon_i$)-差分隐私保护[16].

该性质表明,如果一个差分隐私保护算法序列中所有算法处理的数据集彼此不相交,那么该算法序列构成的组合算法提供的隐私保护水平取决于算法序列中的保护水平最差者,即预算最大者. 该性质也称为并行组合性.

基于差分隐私可组合性的特性,本文实现了深度差分隐私算法. 该算法是在深度学习模型参数最小化损失函数过程中结合差分隐私理论实现的. 本文选取随机梯度下降算法并在其计算过程中添加噪声,其算法实现过程为:首先,采用随机梯度下降算法随机选取小量训练输入样本,计算每个样本的梯度值 $g_t(x_i) \leftarrow \nabla_{\theta_t} L(\theta_t, x_i)$;其次,判断 g_t 是否满足阈值 C,若不满足则调整 g_t,使其在梯度阈值 C 范围内得到新的梯度值 $\bar{g}_t \leftarrow g_t(x_i)/\max(1, \|g_t(x_i)\|_2/C)$;然后,在梯度值 $\bar{g}_t$ 中添加高斯噪声;最后,将添加噪声梯度朝相反方向前进一步,进行下一次的计算. 除了输出该模型的梯度值外,同时还需要计算隐私成本.

深度差分隐私算法中随机梯度下降的实现过程是：首先，计算每个 batch（小批量样本）梯度值；然后，随机选取一组 batch 构成一个 lot，该值的大小影响随机梯度的下降速度，通过对该组中 batch 的梯度值求和得到 lot 的梯度值并对其添加噪声，其中每个 lot 都服从独立分布，概率为 $q = L/N$，N 是输入数据集大小；最后，计算 lot 的梯度平均值，作为代价函数的梯度值. 该算法中运行时间是通过计算训练迭代期（epochs）实现的，其中每个迭代期由 N/L 个 lot 的计算时间组成. 在深度差分隐私算法中，一个重要的问题是如何计算整个训练过程中的隐私损失，该值反映深度学习模型的隐私保护效果. 基于差分隐私可组合的特性，可以在训练过程中直接计算隐私损失，因为训练的每个步骤通常需要多次使用梯度下降，所以会造成隐私预算的积累，本文主要通过时刻积累（moments accountant，MA）方法实现最小化隐私损失，具体实现算法详见 2.2 节.

给出的伪代码中，损失函数主要是单个参数，在多层神经网络中可以按照单层网络进行计算，可以在每一层设置不同的阈值 C 以及不同的噪声规模.

深度差分隐私算法如下：

深度差分稳私算法

输入：训练样本 $\{x_1, x_2, \cdots, x_N\}$，损失函数 $L(\theta) = \frac{1}{N}\sum_i L(\theta, x_i)$

其他参数：学习率 η，噪声规模 σ，随机样本大小 L，梯度阈值 C

初始化：随机值 θ_0

for $t \in$ [T] 循环

　随机选取样本集 L_t，其样本概率为 L/N

　/* 计算梯度 */

　针对每个 $i \in L_t$，计算 $g_t(x_i) \leftarrow \nabla_{\theta_t} L(\theta_t, x_i)$

　/* 修正梯度 */　$\bar{g}_t \leftarrow g_t(x_i) / \max\left(1, \frac{\| g_t(x_i) \|_2}{C}\right)$

　/* 添加噪声 */　$\tilde{g}_t \leftarrow \frac{1}{S}\sum_i \bar{g}_t(x_i) + N$

　/* 梯度下降 */　$\theta_{t+1} \leftarrow \theta_t - \eta \tilde{g}_t$

输出：梯度值 θ_t，计算隐私损失

2.2　隐私损失计算方法

上述深度差分隐私保护算法实现过程中，需要计算隐私损失度，本文通过 MA 方法计算隐私损失，其定义为

给定 2 个近邻数据集 $d, d' \in D^n$，1 个算法 A，辅助输入 aux，1 个输出 $o \in \mathbb{R}$，则 o 的隐私损失定义如下：

$$c(o; A; \text{aux}; d; d') \triangleq \log_2 \frac{\Pr[A(\text{aux}; d) = o]}{\Pr[A(\text{aux}; d') = o]}$$

隐私损失主要依赖算法中添加的噪声规模. 使用可组合定理可计算隐私损失. 该方法适用于具有随机抽样的高斯机制的差分隐私，并且可以为深度差分稳私算法提供更紧密隐私损失的估计. 根据定理 2，算法 A 满足 (ε, δ)-差分隐私等价于隐私损失随机变量在算法 A 的尾边界[17]. 尾边界是高斯分布的重要信息，但是直接计算该值就会使得边界发散，因此，需要计算隐私损失随机变量的线性对数值，然后，利用该值和标准马尔科夫式求取尾边界，即可计算差分隐私的损失. 本文按照差分隐私的可组合性依次计算深度学习模型训练过程中的层次顺序，不断更新隐私损失的状态，因此，在辅助输入建模时需要使得第 λ 次输入的算法 A_λ 是 $\lambda - 1$ 次算法的累计.

对于给定的算法 A，定义第 λ 次的 $\alpha_A(\lambda; \partial UX, d, d')$ 表示为

$$\alpha_A(\lambda; \partial UX, d, d') \triangleq \log_2 E_{o \sim A(\partial UX, d, d')}[\exp(\lambda_c(o; A, \partial UX, d, d'))]$$

为了保证隐私保护机制，需要给定一个约束

$$\alpha_A(\lambda) = \max_{\partial UX, d, d'} \alpha_A(\lambda; \partial UX, d, d')$$

在本方法中一个主要难点是如何计算每一步中的 $\alpha_{A_t}(\lambda)$ 边界值，在随机样本的高斯机制中，满足以下条件：设 u_0 为 $\mathcal{M}(0, \sigma^2)$ 概率密度函数，u_1 为 $\mathcal{M}(1, \sigma^2)$ 概率密度函数，u 为 2 个函数的混合 $u = (1-q)u_0 + qu_1$，则 $\alpha(\lambda) = \log_2 \max(E_1, E_2)$，其中 $E_1 = \mathrm{E}_{z \sim u_0}[(u_0(z)/u(z))^\lambda]$，$E_2 = \mathrm{E}_{z \sim u}[(u(z)/u_0(z))^\lambda]$，便可计算 $\alpha(\lambda)$. 除此之外，根据定理 1 和定理 2[15] 得出 $\alpha(\lambda)$ 边界范围值为 $\alpha(\lambda) \leqslant q^2\lambda(\lambda+1)/(1-q)\sigma^2 + O(q^3/\sigma^3)$.

定理 1　可组合性[17]　算法 A 由 $A_1, \cdots, A_K$ 组成，则对于任意 λ 有 $\alpha_A(\lambda) \leqslant \sum_{i=1}^{k} \alpha_{A_i}(\lambda)$.

定理 2　尾边界[17]　定义为对于任意 $\varepsilon > 0$，算法 A 满足 (ε, δ)-差分隐私，则 $\delta = \min_\lambda \exp(\alpha_A(\lambda) - \lambda\varepsilon)$.

用时刻积累法计算隐私损失的过程如下：

隐私损失计算

输入: 噪声规模 σ,样本率 q,隐私损失计算顺序 moment_order

计算本次循环之间的全部隐私损失

计算本轮的隐私损失

计算 E_1、E_2

计算 $\alpha(\lambda) = \log_2 \max(E_1, E_2)$

输出: 计算隐私损失

2.3 基于 DCGAN 隐私反馈算法实现

为了评价深度差分隐私算法的保护力度,本文通过 GAN 生成攻击者可能得到的最优结果,以此为基础调整深度差分隐私保护算法参数,实现最大程度的攻击防御. GAN 模型主要由生成器(generative model) G 和判别器(discriminative model) D 构成,在整个训练过程中 G 和 D 为"博弈"双方,生成器 G 捕捉样本数据的分布,判别器是一个二分类器,用于判断输入的结果来自于训练数据(而非生成数据)的概率. G 和 D 一般均为非线性映射函数,可以是多层感知机、卷积神经网络等. 在训练过程中,生成器 G 的目标就是尽量生成与原始数据接近的结果去欺骗判别器 D; 而 D 的目标就是尽量把 G 生成的结果和真实数据区分开来. 这样,G 和 D 形成了一个动态的博弈过程,G 和 D 之间的关系[18]可以定义为

$$\min_G \max_D V(D,C) = E_{x\sim p_{\text{data}}(x)}[\log_2 D(x)] + E_{z\sim p_z(z)}[\log_2(1 - D(G(z)))]$$

在实际应用 GAN 的过程中生成器和判别器达到平衡比较困难,因为 GAN 要交替优化 G 和 D,且实现很好的同步. 但是在实际中通常需要多次训练 D,然后再更新 G,如果没有很好地平衡 G 和 D,那么 G 最后就可能会坍缩到一个鞍点. 为了改进这个缺点,Alec 等[19]提出 DCGAN,该方法首先利用批量规范化(batch normalization)解决初始化问题,然后移除全连接层,提升收敛速度,最后利用步幅卷积和微步幅卷积替代池化层,缩减了空间采样,使得训练过程变得稳定,因此,本文选用 DCGAN.

本文以深度差分隐私算法处理的数据为基础,通过 DCGAN 生成数据,并从中选出最接近真实数据集的攻击结果; 比较攻击结果和原始数据集的相似度,若相似度超过设定阈值,则重新对深度差分隐私模型进行参数调优,直到满足条件为止.

3 实验结果及分析

3.1 实验环境及数据集介绍

在本节中,将通过具体的实验来对基于 DCGAN 反馈的深度差分隐私保护算法的效果进行分析、验证和说明. 实验环境为 Intel(R) Xeon(R) CPU E5-2603 v3@1.6 GHz,8 GB 内存,2 块 TITAN X,Ubuntu 16.04 64 位操作系统,实验利用 TensorFlow1.0 框架以及 bazel0.3.1 编译器,该算法由 python 实现.

实验中使用的数据集为 MNIST(手写数字)[20],该数据集包含 60 000 个训练示例和 10 000 个测试示例,其中训练集的文件中包含了 60 000 个标签内容,每个标签的值为 0~9 的一个数,每个示例均为 28×28 的灰度图像.

实验中使用的深度学习网络结构是深度为 3 的前馈神经网络,其隐含层的节点为 1 000,激活函数为 ReLU 和 softmax 为 10 类具有交叉熵的分类器(对应于 10 位数),输入层通过主成分分析(principal component analysis,PCA)进行特征选取,相似度阈值为≤10%.

3.2 相关实验与结果分析

实验 1 不同隐私预算下的深度差分隐私算法实验

基于上述数据集,本实验设置梯度阈值 C 为 4、PCA 为 60 维. 在差分隐私理论中,隐私预算反应了隐私保护的程度,值越小说明隐私保护程度越高,一般取值范围为 0~10. 通过改变隐私预算 ε、隐私偏差 δ、噪声添加规模 σ 来度量算法的可用性.

本次实验分为 3 组: 第 1 组固定 ε 为 2、δ 为 10^{-5},改变 σ,取值为 1~10,实验结果如图 2 所示. 第 2 组固定 σ 为 4,ε 为 0.1~10,δ 为 10^{-5}~10^{-2},实验结果如图 3 所示. 第 3 组固定 δ 为 10^{-5},ε 分别取 0.50、2.00、8.00,σ 分别取 8、4、2,实验结果如图 4 所示.

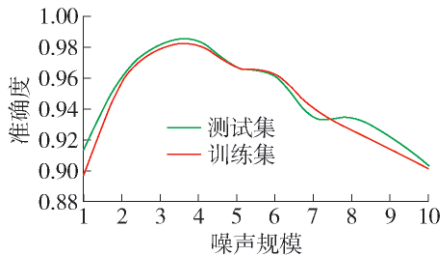


图 2 噪声规模的变化对准确率的影响

Fig. 2 Effect of the change of noise size on accuracy

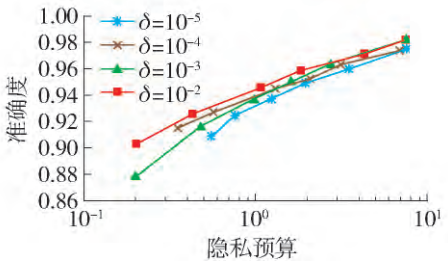


图 3　隐私预算对准确率的影响

Fig. 3　Effect of privacy budget on accuracy

图 2 结果表明，深度模型隐私保护度与 σ 呈倒“U”形曲线关系；图 3 结果表明，ε 与 δ 呈正相关，准确率随着 ε、δ 的增加而增加，改变 ε 对准确率影响较大，而改变 δ 对准确率影响较小. 在深度学习模型训练时，网络的输出 $y(x)$ 能够拟合所有的训练输入 x，本文采用随机梯度下降算法实现寻找最小化权重和偏置的代价函数，该方法通过重复计算梯度 ∇C，然后沿着相反的方向移动，在随机梯度下降过程中添加噪声会影响模型分类效果准确率，在 σ

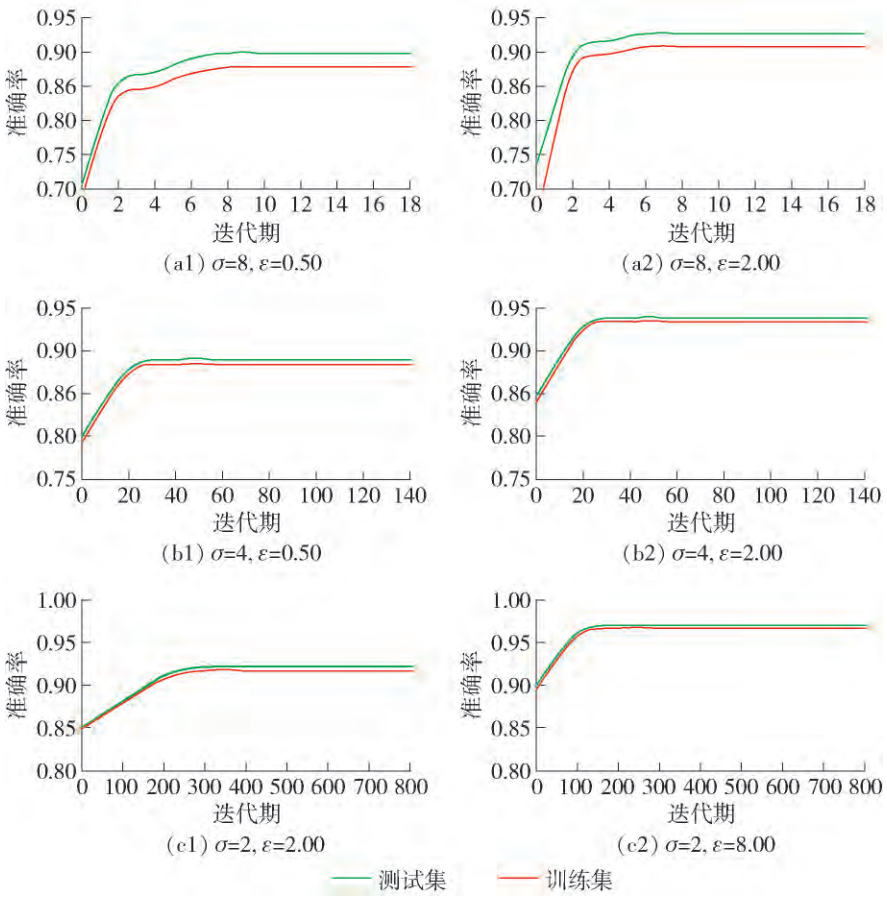


图 4　不同噪声规模、隐私预算对准确率的影响

Fig. 4　Effect of different noise size and privacy budget on accuracy

为 4 时噪声添加规模与模型准确率达到平衡，使得准确率达到最大值.

为了验证 ε 对准确率的影响，实验中分别设置了不同的 σ 和 ε，以分析不同 $\langle\sigma,\varepsilon,\delta\rangle$ 组合对深度网络模型分类准确率的影响. 图 4 结果表明：改变噪声添加量和隐私保护预算值均会影响模型分类准确性，在不同噪声规模下该算法分类准确率分别在 $(8,2,10^{-5})$ $(4,2,10^{-5})$ $(2,8,10^{-5})$ 达到 93%、95%、97%. 当 σ 相同时，ε 与模型分类准确率呈正相关，当 ε 相同时，σ 与模型分类准确率呈负相关.

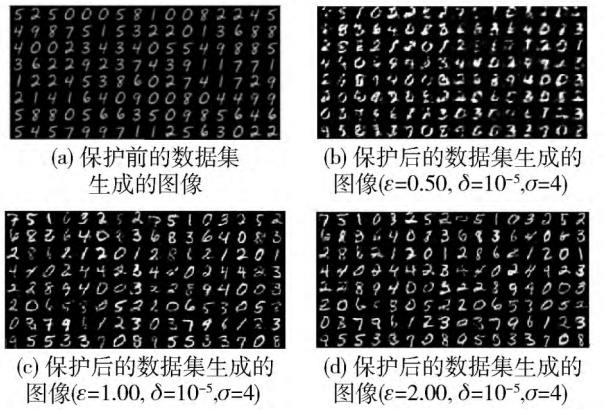
根据图 4 实验结果可知：随着差分隐私预算的增加，噪声添加减少，训练迭代期增长. ε 取值 2，δ 取值 10^{-5}，σ 取值 4 时，深度差分隐私保护模型分类准确率达到 97.00%，与无隐私保护的深度 99.23% [21] 相比，其分类准确率有所下降，但模型可用性仍较高，且选取的参数可取得隐私保护度和数据可用性之间的平衡.

实验 2　隐私保护效果比较实验

本实验主要是利用 DCGAN 生成数据对深度差分隐私算法的有效性进行评价. 其中判别器的输入

来源: 一是深度差分隐私算法处理后的数据集,二由生成器生成的图片,判别器的损失函数 d_loss 由 d_loss_fake 和 d_loss_real 组成,其中 d_loss_real 是深度差分隐私算法处理后的数据集输入到判别器中的结果和预期为 1 的结果之间的交叉熵,d_loss_fake 是生成器生成的图片输入到判别器中的结果和预期为 0 的结果之间的交叉熵,当生成器与判别器达到平衡时,DCGAN 训练结束.

本次实验对不同隐私预算下的效果进行比较,其中第 1 组实验参数设置为: ε 取值 0.50,δ 取值 10^{-5},σ 取值 4; 第 2 组验参数设置为: ε 取值 1,δ 取值10^{-5},噪声添加规模 σ 取值 4; 第 3 组验参数设置为: ε 取值 2,δ 取值10^{-5},σ 取值 4. 其中图 5(a) 为生成器利用原始数据集生成的数据,图 5(b) ~(d) 分别为生成器利用基于 DCGAN 反馈的深度差分隐私保护方法依据第 1、2、3 组设置参数生成的数据.



(a) 保护前的数据集生成的图像 (b) 保护后的数据集生成的图像(ε=0.50, δ=10^{-5},σ=4)

(c) 保护后的数据集生成的图像(ε=1.00, δ=10^{-5},σ=4) (d) 保护后的数据集生成的图像(ε=2.00, δ=10^{-5},σ=4)

图 5 深度差分隐私算法评价

Fig. 5 Evaluation of deep difference privacy algorithm

对比图 5(a) 、5(b) 、5(c) 和 5(d) 可知,当 δ 取值10^{-5},σ 取值 4 时,ε 取值分别为 0.50、1.00、2.00,对应的深度差分隐私保护模型分类准确率分别达到 88.70%、89.00%、89.33%,其隐私预算的大小与模型准确率呈正相关; 从 DCGAN 生成的图像质量来看,经过深度差分隐私算法保护的数据集比利用真实数据集生成的图像辨识度下降,且与 ε 呈正相关. 在 ε 取值 2,δ 取值10^{-5},σ 取值 4 时,深度差分隐私保护模型分类准确率达到 89.33%,满足相似度阈值≤10%,与无隐私保护的深度学习模型 99.13% [11] 相比,其分类准确率有所下降,但模型可用性仍较高.

4 结论

1) 针对深度学习模型在应用中攻击者可利用 DCGAN 等技术还原出训练集数据的问题,本文提出了一个基于 DCGAN 反馈的深度差分隐私保护方法,实现了深度保护训练数据集中用户的敏感信息的目的.

2) 该方法在深度网络参数优化阶段,利用差分隐私与高斯分布可组合特点,在计算过程中融合差分隐私思想,并添加噪声数据; 利用 DCGAN 生成攻击者可能得到的最优结果,通过对比攻击结果和原始数据间差别反馈调节深度差分隐私模型参数实现训练数据集可用性与隐私保护度的平衡.

3) 该方法在实现隐私保护的同时具有更良好的数据可用性,对训练数据集的保护和评价起到一定的帮助作用,数据分析师可以借助生成的数据集进行实验或其他工作,在保证正确率的情况下可以有效地减少信息的泄露.

4) 研究如何在深度学习模型训练过程中保护用户敏感信息将具有广阔的研究前景和实用价值. 本文主要是根据用户的意愿设置参数,通过调节深度差分隐私模型实现保护用户敏感信息的目的,后续工作应围绕如何自适应调节参数使得基于 DCGAN 反馈的深度差分隐私保护方法可满足不同用户的需求.

参考文献:

[1] 周飞燕, 金林鹏, 董军. 卷积神经网络研究综述 [J]. 计算机学报, 2017, 40(7): 1-23.
ZHOU F Y, JIN L P, DONG J. Review of convolutional neural network Chinese [J]. Journal of Computers, 2017, 40(7): 1-23. (in Chinese)

[2] 金连文, 钟卓耀, 杨钊, 等. 深度学习在手写汉字识别中的应用综述 [J]. 自动化学报, 2016, 42(8): 1125-1141.
JIN L W, ZHONG Z Y, YANG Z, et al. Applications of deep learning for handwritten Chinese character recognition: a review [J]. Acta Automatica Sinica, 2016, 42(8): 1125-1141. (in Chinese)

[3] PIERANGELA S. Protecting respondents´ identities in microdata release [J]. IEEE Transactions on Knowledge and Data Engineering, 2001, 13(6): 1010-1027.

[4] ASHWIN W, DANIEL K, JOHANNES G, et al. L-diversity: privacy beyond *k*-anonymity [J]. IEEE Transactions on Knowledge Discovery and Data Mining, 2007, 1(1): 1-5.

[5] JUSTIN B, VITALY S. The cost of privacy: destruction of data-mining utility in anonymized data publishing [C] // ACM SIGKDD International Conference on Knowledge

Discovery and Data Mining. New York: ACM, 2008: 70-78.

[6] REZA S, VITALY S. Privacy-preserving deep learning [C] // ACM SIGSAC Conference on Computer and Communications Security. New York: ACM, 2015: 1310-1321.

[7] MATT F, SOMESH J, THOMAS R. Model inversion attacks that exploit confidence information and basic countermeasures [C] // ACM SIGSAC Conference on Computer and Communications Security. New York: ACM, 2015: 1322-1333.

[8] IAN J G, JEAN P, MEHDI M, et al. Generative adversarial nets [J]. Advances in Neural Information Processing Systems, 2014(3): 2672-2680.

[9] NHAT H P, WANG Y, WU X T, et al. Differential privacy preservation for deep auto-encoders: an application of human behavior prediction [C] // Thirtieth AAAI Conference on Artificial Intelligence. Palo Alto: AAAI Press, 2016: 1309-1316.

[10] NICOLAS P, MARTÍN A, ÚIFAR E, et al. Semi-supervised knowledge transfer for deep learning from private training data [J]. arXiv Preprint arXiv: 1610.05755, 2016.

[11] 张啸剑，孟小峰. 面向数据发布和分析的差分隐私保护 [J]. 计算机学报, 2014, 37(4): 927-949.
ZHANG X J, MENG X F. Differential privacy in data publication and analysis [J]. Journal of Computers, 2014, 37(4): 927-949. (in Chinese)

[12] 熊平，朱天清，王晓峰. 差分隐私保护及其应用 [J]. 计算机学报, 2014, 37(1): 101-122.
XIONG P, ZHU T Q, WANG X F. A survey on differential privacy and applications [J]. Journal of Computers, 2014, 37(1): 101-122. (in Chinese)

[13] ILYA M. Renyi differential privacy [J]. arXiv Preprint arXiv: 1702.07476, 2017.

[14] CYNTHIA D, AARON R. The algorithmic foundations of differential privacy [J]. Foundations and Trends in Theoretical Computer Science, 2014, 9(3/4): 211-407.

[15] JÜRGEN S. Deep learning in neural networks: an overview [J]. Neural Networks, 2015(61): 85-117.

[16] 张啸剑，孟小峰. 基于差分隐私的流式直方图发布方法 [J]. 软件学报, 2016, 27(2): 381-393.
ZHANG X J, MENG X F. Streaming histogram publication method with differential privacy [J]. Journal of Software, 2016, 27(2): 381-393. (in Chinese)

[17] MARTÍN A, ANDY C, IAN J G, et al. Deep learning with differential privacy [C] // Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security. New York: ACM, 2016: 308-318.

[18] 周水庚，李丰，陶宇飞，等. 面向数据库应用的隐私保护研究综述 [J]. 计算机学报, 2009, 32(5): 847-861.
ZHOU S G, LI F, TAO Y F, et al. Privacy preservation in database applications: a survey [J]. Journal of Computers, 2009, 32(5): 847-861. (in Cninese)

[19] ALEC R, LUKE M, SOUMMITH C. Unsupervised representation learning with deep convolutional generative adversarial networks [J]. Computer Science, 2015(4): 1-16.

[20] YANN L C. Courant Institute NYU, MNIST datasets [DB/OL]. [1998-11-01]. http://yann.lecun.com/exdb/mnist/.

[21] PAPERNOT N, MCDANIEL P, SINHA A, et al. Towards the science of security and privacy in machine learning [J]. arXiv Preprint arXiv: 1611.03814, 2016.

（责任编辑　梁　洁）