

基于生成对抗网络的文本序列数据集脱敏

张煜, 吕锡香, 邹宇聪, 李一戈

(西安电子科技大学网络与信息安全学院, 陕西 西安 710071)

摘要: 基于生成对抗网络和差分隐私提出一种文本序列数据集脱敏模型, 即差分隐私文本序列生成网络(DP-SeqGAN)。DP-SeqGAN通过生成对抗网络自动提取数据集的重要特征并生成与原数据分布接近的新数据集, 基于差分隐私对模型做随机加扰以提高生成数据集的隐私性, 并进一步降低鉴别器过拟合。DP-SeqGAN具有直观通用性, 无须对具体数据集设计针对性脱敏规则和对模型做适应性调整。实验表明, 数据集经DP-SeqGAN脱敏后其隐私性和可用性明显提升, 成员推断攻击成功率明显降低。

关键词: 隐私保护; 数据脱敏; 生成对抗网络; 差分隐私

中图分类号: TP309.2

文献标识码: A

doi: 10.11959/j.issn.2096-109x.2020046

Differentially private sequence generative adversarial networks for data privacy masking

ZHANG Yu, LYU Xixiang, ZOU Yucong, LI Yige

School of Cyber Engineering, Xidian University, Xi'an 710071, China

Abstract: Based on generative adversary networks and the differential privacy mechanism, a differentially private sequence generative adversarial net (DP-SeqGAN) was proposed, with which the privacy of text sequence data sets can be filtered out. DP-SeqGAN can be used to automatically extract important features of a data set and then generate a new data set which was close to the original one in terms of data distributions. Based on differential privacy, randomness is introduced to the model, which improves the privacy of the generated data set and further reduces the over fitting of the discriminator. The proposed DP-SeqGAN was universal, so there is no need to adjust the model adaptively for datasets or design complex masking rules against dataset characters. The experiments show that the privacy and usability of a sequence data set are both improved significantly after it is processed by the

收稿日期: 2020-03-12; 修回日期: 2020-04-09

通信作者: 吕锡香, xxlv@mail.xidian.edu.cn

基金项目: 信息保障重点实验室基金(KJ-17-108); 陕西省重点研发计划(2019ZDLGY12-08); 国家重点研发计划(2018YFB0804105)

Foundation Items: The Foundation of Science and Technology on Information Assurance Laboratory (KJ-17-108), The Key Research and Development Project of Shaanxi Province, China (2019ZDLGY12-08), The National Key R&D Program of China (2018YFB0804105)

论文引用格式: 张煜, 吕锡香, 邹宇聪, 等. 基于生成对抗网络的文本序列数据集脱敏[J]. 网络与信息安全学报, 2020, 6(4): 109-119.

ZHANG Y, LYU X X, ZOU Y C, et al. Differentially private sequence generative adversarial networks for data privacy masking[J]. Chinese Journal of Network and Information Security, 2020, 6(4): 109-119.

DP-SeqGAN model, and DP-SeqGAN can greatly reduce the success rate of member inference attacks against the generated data set.

Key words: privacy preserving, data privacy masking, generative adversarial network, differential privacy

1 引言

近年来,深度学习技术有了突破性的进展,在各领域表现出了明显优势。大数据、算法设计以及高性能计算支撑这一切实现的基础,而数据是前提也是重中之重,数据和特征决定了机器学习上限,模型和算法的选择及优化只是在逐步接近这个上限。然而,敏感数据泄露是近年来全球普遍存在和受关注的安全事件,对隐私问题的担忧阻碍着数据的开放、共享和融合,不利于数据价值的充分发挥,也是造成数据孤岛、数据割据的主要原因。因此,设计有效的数据脱敏方法以防范隐私泄露风险,对于打通数据壁垒,消除数据孤岛现象具有积极意义,利于数据的开放、共享和融合,进而促进机器学习的发展。

传统的数据脱敏技术主要通过隐私属性的替换或模糊来实现^[1-3],这类技术在实际应用中存在的问题是:隐私属性的定义和定位难度大;不能抵御重识别攻击^[4];脱敏后数据的可用性有较大损失,表现为下游模型性能严重下降。基于同态加密、安全多方计算等密码学方法是机器学习相关应用中数据隐私保护的重要途径^[5-12],但计算或通信成本高,可能造成相关应用系统性能大幅度下降。面向高效数据脱敏需求,Triastcyn^[4]和 Park^[13]等分别提出了基于生成对抗网络(GAN, generative adversarial network)的适用于图像或表格类结构型数据集的数据合成模型,这种模型能够生成接近于原数据分布的去隐私数据,对图像或表格型数据的脱敏效果表现良好。然而,这种生成模型在文本序列数据上的表现不好^[14],Park等^[13]也提出了隐私文本序列生成模型的开放问题。因此,如何利用生成对抗网络生成高质量的文本序列数据集并且满足数据可用性和隐私性需求,值得研究。

本文针对文本序列数据的隐私保护需求,提出了基于GAN和差分隐私的文本序列数据集脱敏模型DP-SeqGAN,能够在保持较高数据可用性

的同时有效保护隐私,主要优势表现在以下3方面:①具有直观通用性,即无须针对不同文本序列数据集进行适应性的调整;②生成数据的可用性高,用DP-SeqGAN模型所生成的数据训练RNN和CNN模型,其分类准确率相比用原始数据训练的模型分别由0.788和0.914提高到0.914和0.927;③降低成员推断攻击成功率,相比SeqGAN^[15],攻击准确率由0.73和0.61降为0.51和0.50。

2 国内外研究现状

2.1 面向表格型数据的典型脱敏方法

典型的数据脱敏方法主要面向表格型数据,如 k -匿名^[1]、 l -多样性^[2]、 t -closeness^[3]等,这些经典方法通过对数据的隐私属性进行替换或模糊,在数据发布环节实现一定程度的隐私保护。 k -匿名技术通过对敏感属性的泛化处理使每条记录至少与表格中其他 $k-1$ 条记录具有相同的准标识符属性值,从而减少链接攻击所导致的隐私泄露, k -匿名容易遭受一致性攻击与背景知识攻击^[16]。 l -多样性隐私保护技术要求数据表中每个等价类中至少有 l 个可以代替的敏感属性的值,从而使攻击者推断出目标个体敏感信息的概率至多为 $\frac{1}{l}$,该技术不能抵御同质性攻击^[17]。 t -closeness技术要求表中每个等价类中的属性分布和整个表中的属性分布之间的距离不超过门限 t ,但依然不能抵御背景知识攻击。

这类适用于表格数据的匿名化方法不适用于文本序列数据集。一方面,对文本序列数据来说,隐私属性的定义和定位难度大,在不理解完整句、段、篇的情况下,很难找到其中的敏感信息,如个人的喜好信息等;另一方面,对文本序列进行隐私属性模糊、替换等泛化处理会导致脱敏后数据的可用性大大损失。

2.2 面向深度学习的数据隐私保护

利用同态加密能够使深度学习模型在密态数据集上进行训练和预测,数据以密文的形式投放到

开发、共享等下游数据应用。因此, 同态加密是隐私保护最直接的手段。2013年, Graepel等^[10]提出在机器学习算法中使用同态加密并致力于寻找能够在加密数据上训练的学习算法。2015年, Aslett等^[11]提出了可在同态加密数据上实现训练和预测的方法, 适用于朴素贝叶斯分类器和随机森林。Gilad-Bachrach等^[5]基于同态加密方案YASHE^[18]提出CryptoNets, 首次将神经网络用于加密数据的推理, CryptoNets不支持在加密数据上训练模型, 而主要关注基于训练好的CNN模型对密文数据实现预测。Hesamifard等^[6]提出的CryptoDL, 利用Leveled同态加密算法并对激活函数做低次多项式逼近, 提升了模型推理效率, CryptoDL能在加密数据上实现训练和预测。这方面的代表性成果包括TAPAS^[7]和FHE-DiNN^[8], 在预测效率上优于基于Leveled同态加密方案, 并且支持对单个样本的预测。目前, 基于同态加密的隐私保护深度学习面临的最大问题是计算复杂度高, 同态加密能使进程至少慢一个数量级^[6]。结合同态加密和加密电路的安全两方计算协议也被用于在深度学习相关应用场景中的数据隐私保护, 如SecureML^[9], 这类方法带来较高的通信和计算代价。

差分隐私是机器学习中数据隐私保护的主要技术手段之一。基于差分隐私^[19]的隐私保护深度学习主要将训练数据集和模型参数分别对应为数据库和响应, 在满足差分隐私定义的条件下学习模型, 降低了训练数据隐私泄露的风险。根据噪声的添加位置, 这些方法可以分为3种: 梯度级差分隐私、目标函数级差分隐私和标签级差分隐私, 分别是对梯度、目标函数的系数和教师学生模型知识转移阶段的标签添加噪声。梯度级差分隐私方法^[20]的核心是DP-SGD算法^[21], 该算法在批量梯度更新中添加噪声, 限制每个样本对最终模型的影响, 其基于moment accountant算法的累积隐私预算估计方法使该模型的隐私预算相对较小, 因而能提供更好的隐私性。目标函数级差分隐私方法^[20,22]指向经验风险最小化的目标函数表达式中引入随机项, 并保证求解过程满足差分隐私。这类方法要求目标函数是连续、可微的凸函数, 故而具有较大的局限性。标签

级差分隐私方法^[23]在教师学生模型的知识转移阶段对标签引入噪声, 由于学生模型不直接接触数据以及聚合阶段对标签噪声的添加, 这种方法能够在教师模型安全的前提下提供较好的隐私性。针对移动云服务, Wang等^[24]提出基于差分隐私对本地数据进行扰动变换的方法, 同时利用噪声训练方法增加云端深度神经网络的鲁棒性。

2.3 生成对抗网络

自2014年, Goodfellow提出生成对抗网络^[25], 其可应用于计算机视觉、自然语言处理等领域。GAN由生成器和鉴别器构成, 两者进行非合作零和博弈, 交替优化, 生成与原始分布近似的数据集。

自GAN诞生以来, 出现了各种基于GAN的衍生模型进行理论扩展及应用。Arjovsky等^[26]提出的WGAN彻底解决了GAN训练不稳定的问题, 并基本解决崩溃模式现象, 确保了生成样本的多样性。Radford等^[27]提出了深度卷积生成式对抗网(DCGAN), 把有监督学习的CNN与无监督学习的GAN整合, 升级了GAN的架构。Chen等^[28]提出InfoGAN, 结合信息论解释了输入噪声变量的特定变量维数和特定语义之间的关系。最初, GAN的主要应用基本与图像相关, 如图像修改方面的单图像超分辨率^[29]、交互式图像生成^[30]、图像编辑、图像到图像的翻译^[31]等。

标准GAN在处理文本序列这种离散数据时遇到了生成器难以传递梯度更新和鉴别器难以评估非完整序列的问题。为扩展标准GAN的适用范围, 2017年, Yu等^[15]借鉴强化学习的思想, 结合GAN与强化学习的策略梯度算法提出序列生成对抗网络SeqGAN。SeqGAN把整个GAN看作一个强化学习系统, 用策略梯度算法更新生成器的参数, 并借鉴蒙特卡洛树搜索的思想对任意时刻的非完整序列进行评估。

本文提出的文本序列数据集脱敏模型DP-SeqGAN是在SeqGAN的基础上实现的。

3 基于GAN的文本序列数据集脱敏模型

生成对抗网络的诞生为数据脱敏提供了新

的思路,即基于 GAN 生成新的脱敏数据集代替原数据集用以训练下游模型。这种方法能够克服传统数据脱敏方法需要设计针对性规则处理隐私属性的缺陷,这种缺陷主要体现在两方面:① 隐私属性的定义和定位困难;② 规则不具有通用性,即对一个数据集设计的规则并不一定适用于另一个数据集。利用基于 GAN 的数据脱敏模型,工程师只需将待脱敏的数据集输入模型,模型将自动学习数据的特征并“重写”数据,输出即为脱敏后的新数据集。模型对数据的自动学习与刻画,简化了人工操作,使工程师可以简单地将其看作“黑盒”,不需要考虑其内部细节。

实际上,正是 GAN 模型自动学习原数据特征并“重写”输出新数据的原理保证了脱敏后数据的隐私性和可用性。一方面,该方法能够避免传统基于泛化思想的数据脱敏方法对文本序列语法语义的破坏,因而更适用于文本序列数据集;另一方面,“重写”的生成数据与原数据间不存在一对一关系,某些数据的泄露并不会对应到具体实体身份,而且引入差分隐私能够进一步提高生成数据的隐私性。

基于以上理论分析,针对文本序列数据集脱敏需求,本文基于生成对抗网络和差分隐私提出一种文本序列数据集脱敏模型,即差分隐私文本序列生成网络(DP-SeqGAN)。DP-SeqGAN 在训练过程中引入差分隐私,不仅降低了成员推断攻击的成功率,而且在可用性方面有所提高(表现为下游模型性能)。

3.1 模型架构

本文模型借鉴了 SeqGAN 的思想^[15],主要包括生成器(G)和鉴别器(D)两部分,如图 1 所示。生成器用于生成文本序列,将输入的随机噪声加工成文本序列,鉴别器判断其输入样本是原始的训练数据还是由生成的数据,并将判别概率作为奖励(Reward)反馈给 G,用于指导 G 的更新。G 和 D 多次迭代,交替优化,最终生成与原始数据分布接近的数据集。

具体来说,鉴别器是用于文本序列分类的 TextCNN^[32],依次包括嵌入层、卷积层、池化层和 softmax 层,训练集由原始样本和生成器生成的样本组成。对于每个序列,首先将词对应的词向量按行连接成表示序列向量的矩阵,即

$$\mathbf{d}_{1:T} = \mathbf{x}_1 \oplus \mathbf{x}_2 \oplus \cdots \oplus \mathbf{x}_T \quad (1)$$

其中, $\mathbf{x}_i \in R^k$ 是长为 T 的序列中词语 i 对应的 k 维词向量, $\oplus$ 为按行连接操作, $\mathbf{d}_{1:T} \in R^{T \times k}$ 是得到的序列向量矩阵。然后使用不同的卷积核对序列向量矩阵 $\mathbf{d}_{1:T}$ 做卷积操作,提取序列特征,具体如式(2)所示。

$$c_i = \rho(\omega \otimes \mathbf{d}_{i:i+l-1} + b) \quad (2)$$

其中, $\omega \in R^{l \times k}$ 为卷积核, $\otimes$ 为对应位置乘积的和, b 为偏置, ρ 为非线性激活函数。最后将池化后的特征输入全连接 softmax 层,得到输入序列类别为“真实”的概率。

生成器选用长短期记忆(LSTM, long short-term memory)网络^[33],将前一时刻的隐状态和当前时刻的词向量映射为当前时刻的隐状态,

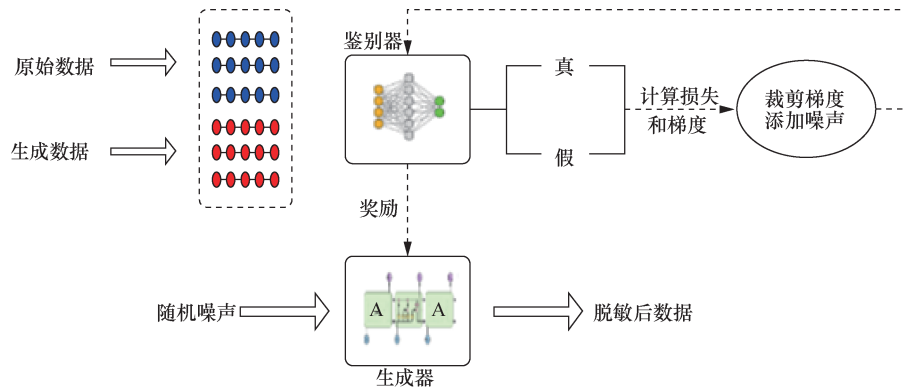


图 1 文本序列数据集脱敏模型框架
Figure 1 Framework of data privacy masking for sequence data sets

可表示为

$$h_t = g(h_{t-1}, x_t) \quad (3)$$

其中, g 是 LSTM 模型等效函数, h_t 为 t 时刻的隐状态。 h_t 后接 softmax 输出层, 将隐状态映射为输出词的概率分布, 即

$$p(y_t | x_1, \dots, x_t) = z(h_t) = \text{softmax}(c + Vh_t) \quad (4)$$

其中, V 为权重矩阵, c 为偏置。

利用蒙特卡洛搜索将 D 回传的判别概率作为奖励, 通过策略梯度指导 G 的更新, 最后 G 和 D 相互博弈, 循环交替地分别优化 G 和 D 来训练所需要的生成式网络与判别式网络, 直到到达纳什均衡点, 可生成高质量的文本序列。为了提高生成数据的隐私性和在有监督任务下的可用性, 本文在此基础上对上述网络的训练过程做了以下调整: ① 按类别生成带标签的样本; ② 基于 DP-SGD (differential privacy stochastic gradient descent) 的思想^[21], 在鉴别器的优化过程中进行梯度裁剪并加噪声。

3.2 生成带标签数据

基于带标签数据训练模型的有监督学习目前仍是机器学习的主流, 然而现有的生成对抗网络是不能生成带标签数据的, 这是因为它只是从所学到的数据分布中采样, 整个过程并没有用到标签数据。针对生成带标签数据的需求, Mirza 等^[34]提出条件生成对抗网络, 通过将标签作为辅助参数添加到生成器来生成对应类别数据, 然而生成器的回归问题造成的任意大错误输出会大大降低这一网络在实际中的可靠性。现有的针对文本序列的条件生成对抗模型^[35]需要对原网络做复杂

的调整, 并且在数据类别少时生成质量不高, 这是因为学习出的鉴别器过于强大, 反馈到生成器的信息缺乏指导价值, 从而导致生成数据的质量降低。

针对文本序列生成, 本文提出一个生成带标签序列的直接有效的方法, 即根据数据的类别对训练数据集做出划分, 按照此划分分别输入模型, 则生成的数据可以继承输入数据的类别, 最后将这些输出数据汇总并打乱即得到带标签的生成数据集, 具体过程如图 2 所示。该方法不需要修改模型输入、结构或者优化函数, 避免了这些操作所带来的不确定性。更重要的是, 这种方法避免了在类别较少的情况下学习出的鉴别器过于强大的问题, 促使生成器接收更多对生成数据有用的指导信息, 最终生成与原数据分布接近的数据, 数据可用性也与原始数据接近。

生成的带标签文本序列数据可以直接用于数据集的扩充, 通过添加不同比例的生成数据, 可以让模型学习到更多更有效的特征, 并且减小了噪声样本的影响, 达到减轻过拟合的效果。这里直接用生成数据代替原始数据完成下游模型的训练, 并且通过差分隐私的引入, 保证数据的隐私性。

3.3 训练过程差分隐私化

成员推断^[36]等攻击已经具备从机器学习模型中重构或恢复出训练样本的能力, GAN 是模型的组合, 自然受到这些攻击的威胁。此外, 具有高模型复杂度的 GAN 会记住某些训练样本, 造成学习到的分布集中在这些样本上, 进而增加了受到攻击的风险。

差分隐私可以在很大程度上缓解这种影响,

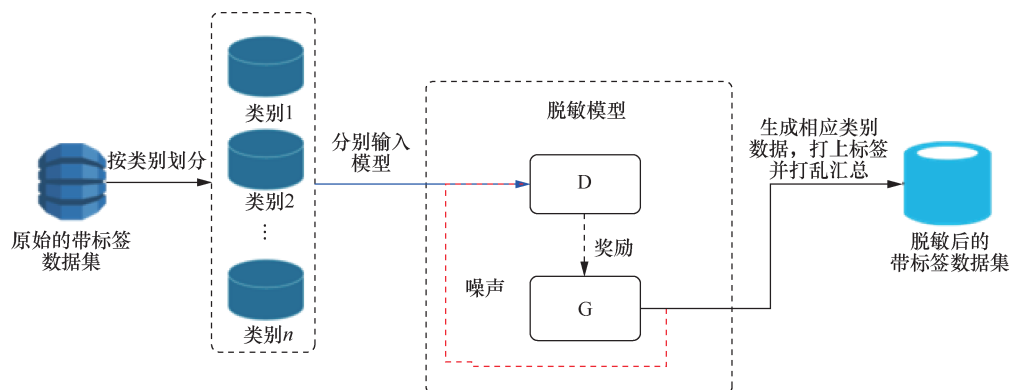


图2 生成带标签文本
Figure 2 Generate tagged texts

根据其定义,如果训练过程是差分隐私的,则训练数据集包含与不包含某一特定数据训练出相同模型的概率是接近的。

为了引入差分隐私,需要“隐私化”鉴别器的训练,这是因为模型的对抗性训练过程中只有鉴别器接触真实数据,生成器根据鉴别器的反馈信息调整自己的参数,由差分隐私的后处理定理^[19],整个网络满足差分隐私保证。由于鉴别器选用的是 CNN,可以基于 DP-SGD 的思想来差分隐私地训练鉴别器,即在随机梯度更新的过程中添加噪声。具体地说,需要做以下两步修改。

(1) 限制每个梯度的敏感度,即限制每个训练样本梯度的大小,这可以通过裁剪在每个训练样本上计算出的梯度 $g(x_i)$ 的 L2-范数来完成,即

$$\bar{g}(x_i) = \frac{g(x_i)}{\max\left(1, \frac{\|g(x_i)\|_2}{C}\right)} \quad (5)$$

其中, $\|g(x_i)\|_2$ 为梯度 $g(x_i)$ 的 L2-范数, C 为裁剪阈值,小于 C 的梯度保持不变,大于 C 的梯度范数缩小到 C ,这样可以限制每个训练样本对模型参数产生的影响。

通过裁剪梯度,控制每个训练样本梯度的大小,一方面限制了每个梯度的敏感度,进而降低了每个训练样本对模型参数产生的影响,缓解了模型对训练样本的记忆;另一方面防止了梯度爆炸,降低了过拟合,使模型能更好地收敛。

(2) 对算法引入随机性,主要是通过给裁剪好的梯度添加从高斯分布中采样的随机噪声实现的,即

$$\tilde{g} = \frac{1}{L} \left(\sum_i \bar{g}(x_i) + N(0, \sigma^2 C^2) \right) \quad (6)$$

其中, L 为一批数据的数量, σ 为噪声大小,该噪声具有与梯度裁剪成正比的方差,添加的噪声越多,隐私保证越好。这样做降低了通过比较随机梯度下降过程的参数更新,来确定训练集中包含特定样本的可能性。

随机性的引入使分辨模型的哪些行为来自随机性、哪些来自训练数据变得困难。没有随机性,本文关心的问题是当在特定数据集上训练时模型会选择哪些参数。有了这种随机性,问题转变为

当在特定数据集上训练时模型在这组的参数中选择的可能性。

通过以上两步操作,差分隐私化鉴别器的训练,确保在添加、删除或更改训练集中的单个训练样本的情况下学习出任何特定参数集的概率保持大致相同。换句话说,如果单个训练样本不影响学习的结果,则该样本中包含的信息将无法被记忆,进而贡献该样本的个人隐私将会得到保护。

4 实验与性能评估

本节详细阐述 DP-SeqGAN 模型的具体实现以及性能评估。本模型的目标是生成隐私文本序列,因此对模型的性能评估可以转换为对生成数据的评估,即评估生成数据是否可以在保护隐私的前提下代替原始数据集完成相关的机器学习任务,具体表现为生成数据的隐私性与可用性两个指标。

实验是在 Intel Core i5-7500 3.40 GHz CPU、8 GB RAM 的硬件条件下,以 Python3.6 环境进行模型训练和性能测试,基于 TensorFlow 框架在 VScode 集成环境中开发模型,在交互式开发环境 Jupyter Notebook 中利用机器学习算法库 Scikit-learn 和数据分析软件库 Pandas 进行可用性与隐私性的测试。

DP-SeqGAN 模型训练和测试所用的数据来自 YouTube 恶意评论检测数据集^[37],该数据集由 2 000 余热门音乐视频的正常评论(ham)和恶意评论(spam)组成,实验中将这数据汇总、打乱并划分为 3 部分,分别是用于训练生成模型的训练集 S_1 、成员推断攻击过程中攻击模型的测试集 S_2 ,以及影子模型的测试集 S_3 (S_3 也作为衡量下游模型性能的留出集)。

DP-SeqGAN 模型的训练基于 Python 和 TensorFlow 框架,并在 SeqGAN 的基础上借鉴 TF-Privacy 的 optimizers.dp_optimizer 模块实现差分隐私优化器,将 SeqGAN 的 AdamOptimizer 修改为 DPAdamGaussianOptimizer。训练中将一批样本测值与标签的损失向量输入差分隐私优化器中,然后进行梯度裁剪和添加噪声,求其平均值用于之后的参数更新。实验中反复调试,最优的梯度裁剪阈值 L2_norm_clip 取为 1,噪声控制

参数 `noise_multiplier` 取为 0.001。

如前所述, 评估生成数据集的性能需要从隐私性和可用性两个方面进行测试, 测试中以 SeqGAN 直接生成的数据作为对照组。

4.1 隐私性评估

本节从差分隐私和成员推断攻击两方面评估生成数据的隐私性。

4.1.1 基于差分隐私的隐私性评估

Dwork^[19]定义的 (ϵ, δ) 差分隐私是指, 给定随机算法 M , P 为算法 M 所有可能输出构成的集合, 对于任意的一对相邻数据集 D 和 D' , 以及 P 的任意子集 S , 若算法 M 满足

$$\Pr[M(D) \in S] \leq \Pr[M(D') \in S] \cdot e^\epsilon + \delta \quad (7)$$

则称算法 M 满足 (ϵ, δ) -差分隐私。其中, ϵ 和 δ 非负, ϵ 称为隐私预算, 当数据集有 n 个样本时, δ 要小于 $\frac{1}{n}$ 。 ϵ 和 δ 越小, 两个概率越相近, 隐私保护越好。然而, 对于如何选择适当的隐私预算 ϵ , 并未给出相关的指导性工作, 实际中也不存在一个共识的参考值。

差分隐私能够限制单个数据对模型的影响, 即这条数据的隐私就得到了隐藏。本节基于瑞利差分隐私^[38]进行隐私分析, 瑞利差分隐私是纯差分隐私的扩展, 特别适合分析添加高斯噪声的差分隐私保护。

实验中, 在 $\delta=1 \times 10^{-5}$ 时, 测得隐私预算 ϵ 约 4×10^8 , 根据以往的经验, 这个数字是很大的。然而, 对于隐私预算并没有可借鉴的参考阈值。所以, 实际中需要通过测试隐私性有无相对性提高来评估隐私保护的程。同样的情况出现在其他论文中, 如在文献[39]中, 也是文本序列任务模型, Carlini 等对他们的模型梯度只添加很小的噪声, 其隐私预算 ϵ 很大 (与本文实验的 ϵ 在同一数量级), 以至于理论上无法提供有意义的隐私保证, 但他们测量到的隐私泄露并不多。这证明了差分隐私所能提供的隐私保证或隐私预算 ϵ 的上限与用现有实际攻击测得的隐私损失之间存在很大的差距。

为了测试在实际攻击中 DP-SeqGAN 对隐私性有无相对提升, 本文进行了基于成员推断攻击的隐私性评估测试。

4.1.2 基于成员推断攻击的隐私性评估

成员推断攻击^[36]是对机器学习模型的一种新型攻击, 是训练数据隐私性的主要威胁, 攻击者能够判断一个样本是否属于模型的训练集, 进而重构训练数据。成员推断攻击的攻击目标是分类模型, 攻击者向分类模型输入给定样本, 然后判定给定样本是否在模型训练集中。在 DP-SeqGAN 模型中, 只有鉴别器是分类器, 而且只有鉴别器接触训练数据, 对它实施攻击就可以评测训练数据的隐私泄露程度。另外, 本实验中, 鉴别器的训练数据包括真实样本和生成样本, 因此, 成员推断攻击转化为向 DP-SeqGAN 的鉴别器输入给定样本, 然后判定给定样本是属于鉴别器训练集中的真实样本还是生成样本。

假设攻击者具有以下能力: ① 能够得到任意多目标攻击模型的输出; ② 了解攻击目标模型的算法和架构。基于这样的假设, 对每一类别的数据, 本文利用 DP-SeqGAN 生成数据构成影子模型训练集, 训练多个影子模型来模仿鉴别器的行为。训练好影子模型之后, 将影子模型的训练集中的样本输入影子模型, 得到预测向量, 对相应的预测向量 PV 打标签 “In”, 这里的 “In” 表示输入样本在影子模型的训练集中; 将留出的测试集 S_3 中的样本输入影子模型, 得到预测向量, 对相应的预测向量 PV 打标签 “Out”, 这里的 “Out” 表示输入样本在测试集 S_3 中 (即不在影子模型的训练集中)。所得到的打上标签的 $(PV, \text{In/Out})$ 构成攻击模型的训练集, 因为属于影子模型训练集的样本和不属于影子模型训练集的样本所对应的预测向量 PV 不相同, 训练好的攻击模型能够分辨出这种差异。成员推断攻击正是利用所训练的攻击模型的这种分辨能力对目标模型实施攻击。实施攻击时, 攻击者将一条给定的数据输入目标模型得到预测向量 PV , 再将预测向量 PV 输入攻击模型得到结果 “In” 或 “Out”, 从而推断出输入的这条数据是否属于目标模型的训练集。

具体流程如下。

(1) 记 T 为训练好的攻击者想要攻击的目标模型, 这里为 D , 由于本文的数据是按类生成的,

所以攻击也是按类进行的。

(2) 假设攻击者足够强大, 已经通过攻击下游任务模型恢复出生成数据。

(3) 用这些生成数据训练多个影子模型。

(4) 对每个训练好的影子模型, 输入影子训练样本或影子测试样本 r , 由影子模型的一对输入和输出获得攻击模型的一个训练样本 $(D(r), \text{In/Out})$, 其中, $D(r)$ 为鉴别器输出的样本为真的预测概率, 即预测向量 $\mathbf{PV}$, In 代表影子训练样本, Out 代表影子测试样本, 影子测试样本来自于留出的 S_3 。

(5) 将所产生的这些样本 $(D(r), \text{In/Out})$ 汇总成攻击模型训练集。

(6) 用攻击模型训练集训练攻击模型。

(7) 对 D 输入给定样本, 实施攻击。

本次实验中的攻击模型选用 SVM, 攻击模型测试数据集一半来自生成模型的训练集, 一半来自留出的测试集 S_2 , 因此攻击准确率 0.5, 即随机猜测, 意味着攻击失败。

表 1 对比了上述实验在本文提出的 DP-SeqGAN 模型上的攻击成功率和在 SeqGAN 模型上的攻击成功率。显而易见, DP-SeqGAN 明显降低了攻击准确率, 逼近 0.5 的随机猜测概率。

表 1 成员推断攻击成功率
Table 1 Success rate of membership inference attack

模型	Ham (YouTube 正常评论)	Spam (YouTube 恶意评论)
SeqGAN	0.73	0.61
DP-SeqGAN	0.51	0.50

4.2 可用性评估

为测试和评估模型生成数据集的可用性, 即生成数据在下游任务模型上的性能, 本文分别用原始数据、SeqGAN 生成数据和 DP-SeqGAN 生成数据训练分类模型, 在留出的测试集 S_3 上测试三者的分类准确率。对于每个分类模型, 除了训练集不同, 其他操作相同, 包括数据预处理、文本向量化等过程, 实验结果对比如图 3 所示。其中, 绿色为原始数据集, 蓝色为 SeqGAN 生成数据, 灰色为 DP-SeqGAN 生成数据。由图 3 可知, 在 SeqGAN 生成的数据集上训练的模型其分类表现相比用原始数据训练的模型有所降低; 用

DP-SeqGAN 生成数据训练的模型明显优于 SeqGAN 生成的数据; 在 CNN 和广泛应用于文本相关任务的 RNN 上, DP-SeqGAN 生成数据训练的模型明显优于用原始数据训练的模型, 即文本序列数据集经本文提出的 DP-SeqGAN 模型脱敏后, 不但能保护数据的隐私, 而且其可用性有明显提升。

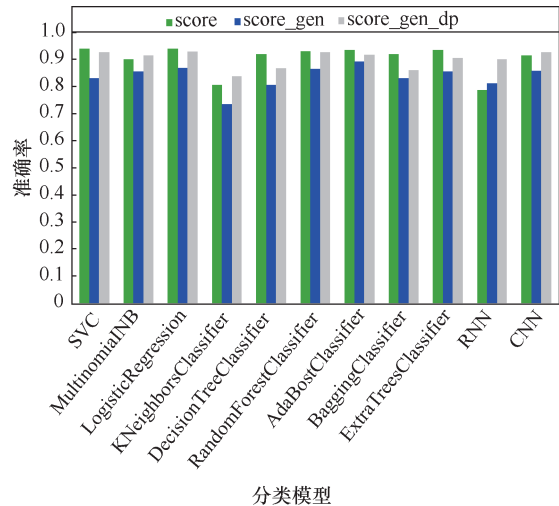


图 3 不同分类模型在 3 个数据集上的分类准确率
Figure 3 Accuracy of different classification models on three data sets

需要特别说明的是, 一方面, 基于 GAN 的数据生成模型 (如 SeqGAN) 生成的是接近于原数据分布的数据, 可用性降低; 另一方面, 引入隐私保护会一定程度减少原数据的信息量, 进而导致脱敏后的数据可用性降低。然而, 针对 DP-SeqGAN 模型的实验表明, 文本序列数据集经 DP-SeqGAN 模型脱敏后其可用性相对原数据集有明显提升, 这是因为加入差分隐私在一定程度上降低了鉴别器的过拟合^[40]和模型的记忆性^[36], 训练集中的一些异常值和未包含分布主要特征的数据会被归为假, 而且鉴别器会进一步将这些信息传递给生成器, 从而生成的是反应主要特征和内在分布的数据, 故而在这些生成数据上训练的下游模型表现出更好的性能。换句话说, 鉴别器加上差分隐私相当于对数据做了预先特征选择, 选择带有主要特征的数据, 并通过生成更多的数据实现主要特征的多次重复。因此, DP-SeqGAN 生成的数据集在 RNN、CNN 上的表现明显超越原始数据, 其根本原因是神经网络模型更容易过

拟合于原始数据集中的一些异常样本,而加上差分隐私的 DP-SeqGAN 有效过滤了原始数据集的异常样本。

在实验中还发现,在 RNN 的训练过程中,用生成数据训练的模型会更早收敛。用原始数据训练的模型约在 6 个 epoch 后收敛,而 DP-SeqGAN 生成数据对应的模型只需要 1 个 epoch 即收敛,

并且 DP-SeqGAN 的生成数据对应的模型更加稳定。RNN 训练过程对比如图 4 所示。

5 结束语

本文结合生成对抗网络与差分隐私机器学习,提出了隐私文本序列数据集脱敏模型 DP-SeqGAN,该模型具有直观通用性,无须针对

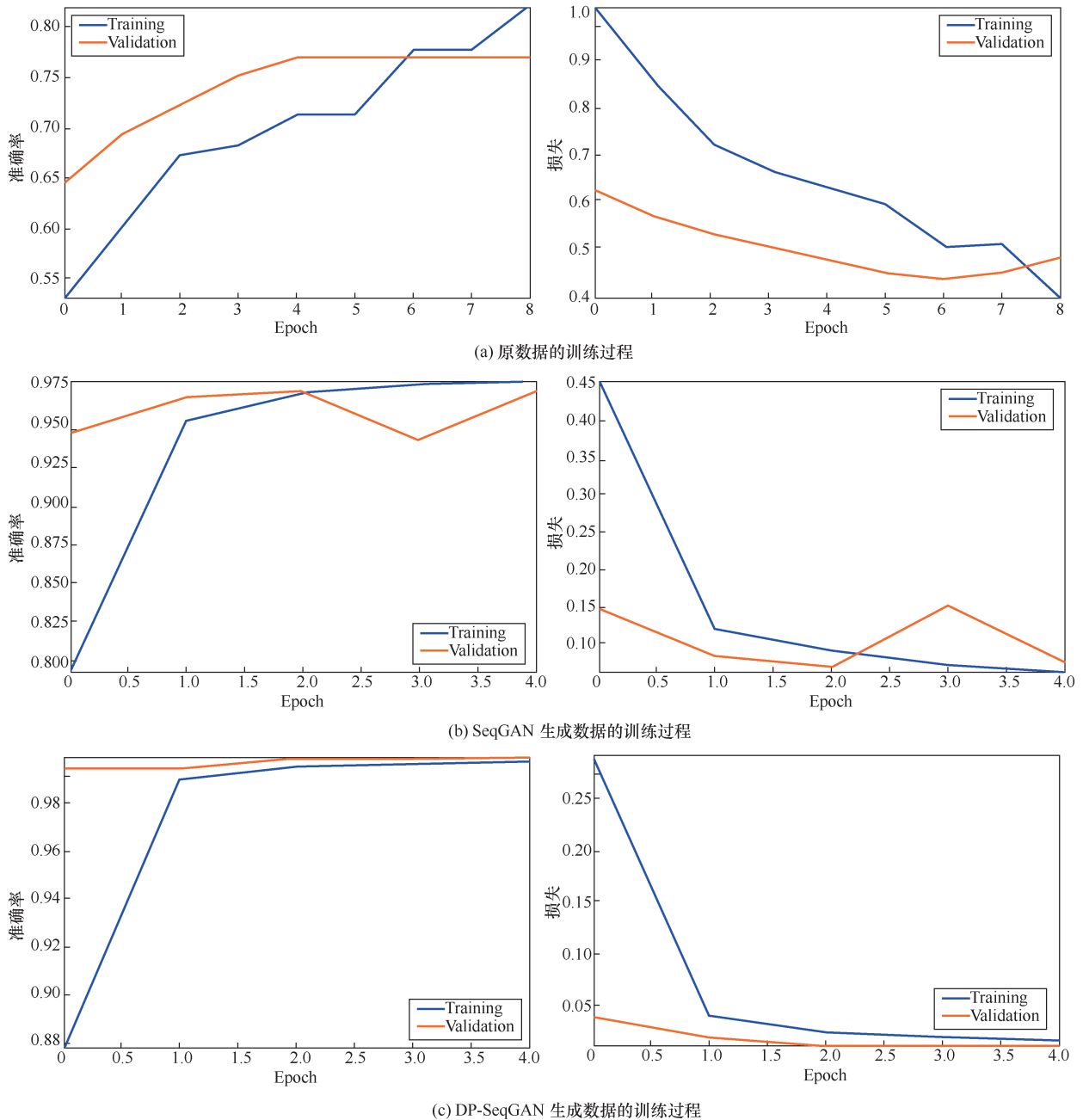


图4 RNN 训练过程对比
Figure 4 Comparison of RNN training process

数据设计脱敏规则和针对数据集对模型进行适应性调整。实验证明,经该模型脱敏后的文本序列数据集在隐私性和可用性上都得到了明显提升。DP-SeqGAN 适用于文本序列这种非结构型的数据,其性能还需在实际应用中进一步验证,也需要结合具体任务类型扩展其应用场景。

参考文献:

- [1] SWEENEY L. *K*-anonymity: a model for protecting privacy[J]. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 2002, 10(5): 557-570.
- [2] MACHANAVAJJHALA A, KIFER D, GEHRKE J, et al. *L*-diversity: privacy beyond *k*-anonymity[J]. ACM Transactions on Knowledge Discovery from Data (TKDD), 2007, 1(1): 3.
- [3] LI N, LI T, VENKATASUBRAMANIAN S. T-closeness: privacy beyond *k*-anonymity and *L*-diversity[C]//IEEE 23rd International Conference on Data Engineering. 2007: 106-115.
- [4] TRIASTCYN A, FALTINGS B. Generating artificial data for private deep learning[C]//Proceedings of the PAL: Privacy-Enhancing Artificial Intelligence and Language Technologies, AAAI Spring Symposium Series. 2019.
- [5] GILAD-BACHRACH R, DOWLIN N, LAINE K, et al. CryptoNets: applying neural networks to encrypted data with high throughput and accuracy[C]//International Conference on Machine Learning. 2016: 201-210.
- [6] HESAMIFARD E, TAKABI H, GHASEMI M. Cryptodl: deep neural networks over encrypted data[J]. arXiv preprint arXiv:1711.05189, 2017.
- [7] SANYAL A, KUSNER M, GASCON A, et al. TAPAS: tricks to accelerate (encrypted) prediction as a service[C]//International Conference on Machine Learning. 2018: 4490-4499.
- [8] BOURSE F, MINELLI M, MINIHOLD M, et al. Fast homomorphic evaluation of deep discretized neural networks[J]. IACR Cryptology ePrint Archive, 2017.
- [9] MOHASSEL P, ZHANG Y. SecureML: a system for scalable privacy-preserving machine learning[C]//2017 IEEE Symposium on Security and Privacy (SP). 2017: 19-38.
- [10] GRAEPEL, THORE, KRISTIN, et al. MI confidential: machine learning on encrypted data[C]//Information Security and Cryptology-ICISC 2012. 2012: 1-21.
- [11] ASLETT L J M, ESPERANÇA P M, HOLMES C. A review of homomorphic encryption and software tools for encrypted statistical machine learning[J]. Stat, 2015, 1050: 26.
- [12] 宋蕾, 马春光, 段广略. 机器学习安全及隐私保护研究进展[J]. 网络与信息安全学报, 2018, 4(8): 1-11.
SONG L, MA C G, DUAN G H. Machine learning security and privacy: a survey[J]. Chinese Journal of Network and Information Security, 2018, 4(8): 1-11.
- [13] PARK N, MOHAMMADI M, GORDE K, et al. Data synthesis based on generative adversarial networks[J]. Proceedings of the VLDB Endowment, 2018, 11(10): 1071-1083.
- [14] REED S, AKATA Z, YAN X, et al. Generative adversarial text to image synthesis[C]//International Conference on Machine Learning. 2016: 1060-1069.
- [15] YU L, ZHANG W, WANG J, et al. SeqGAN: sequence generative adversarial nets with policy gradient[C]//Thirty-First AAAI Conference on Artificial Intelligence. 2017.
- [16] MAYILVELKUMAR P, KARTHIKEYAN M. *L*-diversity on *k*-anonymity with external database for improving privacy preserving data publishing[J]. International Journal of Computer Applications, 2012, 54(14): 7-13.
- [17] WANG Q, XU Z W, QU S Z, et al. An enhanced *k*-anonymity model against homogeneity attack[J]. Journal of Software, 2011: 1945-1952.
- [18] BOS J W, LAUTER K, LOFTUS J, et al. Improved security for a ring-based fully homomorphic encryption scheme[C]//IMA International Conference on Cryptography and Coding. 2013: 45-64.
- [19] DWORK C. Differential privacy: a survey of results[C]//International Conference on Theory and Applications of Models of Computation. 2008: 1-19.
- [20] PHAN N, WANG Y, WU X T, et al. Differential privacy preservation for deep auto-encoders: an application of human behavior prediction[C]//AAAI Conference on Artificial Intelligence. 2016.
- [21] 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. 2016: 308-318.
- [22] CHAUDHURI K, MONTELEONI C. Privacy-preserving logistic regression[C]//Advances in Neural Information Processing Systems. 2009: 289-296.
- [23] PAPERNOT N, ABADI M, ERLINGSSON U, et al. Semi-supervised knowledge transfer for deep learning from private training data[J]. Stat, 2017, 1050: 3.
- [24] WANG J, ZHANG J G, BAO W D, et al. Not just privacy: improving performance of private deep learning in mobile cloud[C]//Proceedings of ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD). 2018.
- [25] GOODFELLOW I. Generative adversarial nets[C]//NIPS. 2014: 2672-2680.
- [26] ARJOVSKY M, CHINTALA S, BOTTOU L. Wasserstein generative adversarial networks[C]//Proceedings of the 34th International Conference on Machine Learning 70. 2017: 214-223.
- [27] RADFORD A, METZ L, CHINTALA S. Unsupervised representation learning with deep convolutional generative adversarial networks[J]. arXiv preprint arXiv:1511.06434, 2015.
- [28] CHEN X, DUAN Y, HOUTHOOFT R, et al. InfoGAN: interpretable representation learning by information maximizing generative adversarial nets[C]//Proceedings of the 2016 Neural Information Processing Systems of Information Technology IMEC. 2016: 2172-2180.

- [29] LEDIG C, THEIS L, HUSZÁR F, et al. Photo-realistic single image super-resolution using a generative adversarial network[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 4681-4690.
- [30] ZHU J Y, KRÄHENBÜHL P, SHECHTMAN E, et al. Generative visual manipulation on the natural image manifold[C]//European Conference on Computer Vision. 2016: 597-613.
- [31] ISOLA P, ZHU J Y, ZHOU T, et al. Image-to-image translation with conditional adversarial networks[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 1125-1134.
- [32] KIM Y. Convolutional neural networks for sentence classification[C]//Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2014: 1746-1751.
- [33] HOCHREITER S, URGEN SCHMIDHUBER J, ELVEZIA C. Long short-term memory[J]. Neural Computation, 1997, 9(8): 1735-1780.
- [34] MIRZA M, OSINDERO S. Conditional generative adversarial Nets[J]. arXiv preprint arXiv:1411.1784, 2014.
- [35] LI Y, PAN Q, WANG S, et al. A generative model for category text generation[J]. Information Sciences, 2018, 450: 301-315.
- [36] SHOKRI R, STRONATI M, SONG C, et al. Membership inference attacks against machine learning models[C]//2017 IEEE Symposium on Security and Privacy (SP). 2017: 3-18.
- [37] YouTube spam collection data set[EB].
- [38] MIRONOV I. RÉNYI. Differential privacy[C]//2017 IEEE 30th Computer Security Foundations Symposium (CSF). 2017: 263-275.
- [39] CARLINI N, LIU C, ERLINGSSON Ú, et al. The secret sharer: evaluating and testing unintended memorization in neural networks[C]//28th USENIX Security Symposium. 2019: 267-284.
- [40] DWORK C, FELDMAN V, HARDT M, et al. The reusable holdout: preserving validity in adaptive data analysis[J]. Science, 2015, 349(6248): 636-638.

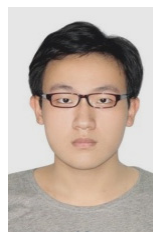
[作者简介]



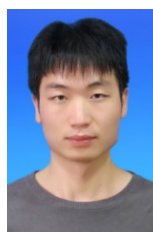
张煜（1995—），男，陕西延安人，西安电子科技大学硕士生，主要研究方向为隐私保护和机器学习。



吕锡香（1978—），女，陕西洛南人，西安电子科技大学教授、博士生导师，主要研究方向为网络与协议安全、机器学习与安全、密码算法与协议。



邹宇聪（1999—），男，湖南桃江人，主要研究方向为隐私保护和机器学习。



李一戈（1995—），男，陕西洛南人，西安电子科技大学博士生，主要研究方向为机器学习与安全。