计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.201200224

基于特征映射的差分隐私保护机器学习方法

陈天荣　凌　捷
广东工业大学计算机学院　广东 510006
(1181113557@qq.com)

摘　要　图像分类中的差分隐私算法在通过添加噪声的方式提高机器学习模型的隐私保护能力的同时,容易造成模型分类准确度的下降。针对以上问题,提出了一种基于特征映射的差分隐私保护机器学习方法,该方法结合预训练神经网络和影子模型训练技术,以差分向量的形式将原数据样本的特征向量映射到高维向量空间,缩短样本在高维向量空间的距离,以减小模型更新造成的隐私信息泄露风险,同时提高机器学习模型的隐私保护能力和分类能力。由 MNIST 和 CIFAR-10 数据集上的实验结果表明,ε 分别等于 0.01 和 0.11 的 ε0-差分隐私的模型的分类准确度分别提高到了 99%和 96%,说明所提方法训练的模型相比 DP-SGD 等现有多种常用差分隐私算法,能在更低隐私预算下保持更强的分类能力;且在两个数据集上针对该模型的推理攻击成功率降低为 10%,其对推理攻击的防御能力相比传统图像分类的 CNN 模型有较大幅度的提升。

关键词:机器学习;差分隐私;图像分类;推理攻击;影子模型
中图法分类号　TP391

Differential Privacy Protection Machine Learning Method Based on Features Mapping

CHEN Tian-rong and LING Jie
School of Computer,Guangdong University of Technology,Guangdong 510006,China

Abstract　The differential privacy algorithm in image classification improves the privacy protection capability of the machine learning model by adding noise,and at the same time easily causes the accuracy of the model classification to decrease. To solve the above problems,a differential privacy protection machine learning method based on features mapping is proposed. This method combines the pre-training neural network and shadow model training technology to map the feature vectors of the original data sample to the high-dimensional vector space in the form of differential vectors,so as to shorten the distance of the sample in the high-dimensional vector space to reduce the leakage of private information caused by model updates,and improve the privacy protection and classification capabilities of the machine learning model. The experimental results on the MNIST and CIFAR-10 datasets show that for the ε-differential privacy model with ε equal to 0.01 and 0.11,the classification accuracy is improved to 99% and 96%,respectively,indicating that compared with DP-SGD and many other commonly used differential privacy algorithms,the model trained by this method can maintain stronger classification capabilities at a lower privacy budget. And the success rate of reasoning attacks against this model on the two data sets is reduced to 10%,which is against inference attacks. Compared with the traditional CNN model of image classification,the defense capability of the CNN model is greatly improved.

Keywords　Machine learning,Differential privacy,Image classification,Inference attack,Shadow model

1　引言

机器学习算法逐渐被应用到各个领域并取得了良好的效果,在机器学习算法被广泛应用的同时,如何保障数据隐私的安全性也成为了一个重要的课题,众包数据、在线学习等新的数据收集模式使数据隐私保护问题变得更加急迫。

针对机器学习隐私安全问题,多种基于机器学习的隐私保护方案被提出,其中差分隐私机器学习算法是主要的解决方案之一。本质上,现有差分隐私算法通过在机器学习的不同阶段添加噪声,来扰动模型的输入或输出以实现隐私保护效果。这种方式破坏了数据的真实性,在提高隐私保护能力的同时,会不可避免地导致模型的分类能力下降[1]。

本文提出了一种基于特征映射的差分隐私保护机器学习方法。与添加噪声的方式不同,该方法中,机器学习模型将学习由训练样本派生的差分信息,以保持样本、梯度等数据的真实性。通过提高模型泛化能力的方式,提高模型的隐私保护

到稿日期:2020-12-25　返修日期:2021-02-20　本文已加入开放科学计划(OSID),请扫描上方二维码获取补充信息。
基金项目:广东省重点领域研发计划项目(2019B010139002);广州市重点领域研发计划项目(202007010004)
This work was supported by the Key Field R&D projects in Guangdong Province of China(2019B010139002) and Key Field R&D projects in Guangzhou(202007010004).
通信作者:凌捷(1150181103@qq.com)

能力。其中,该差分信息是由预训练神经网络结合影子训练技术所获得的特征映射。

本文的贡献如下:

(1)提出了基于特征映射的差分隐私保护机器学习方法,该方法通过将预训练网络与主网络级联的方式,隐藏前者输出端和后者输入端以保证模型安全,并通过特征映射为数据提供隐私安全保障,为隐私保护机器学习算法提供了一个新的方案。

(2)应用 MNIST 和 CIFAR-10 数据集对该模型进行实验和性能评估。实验数据表明,在两个数据集上,隐私预算 ε 分别等于 0.01 和 0.11 的 ε-差分隐私模型,分类准确度分别达到了 99%和 96%。这说明,相比 DP-SGD 等现有多种常用差分隐私算法,本文方法具有更好的隐私保护能力和分类能力。

(3)使用基于差分的推理攻击方法检验本文方法训练的机器学习模型对推理攻击的防御效果。实验数据表明,对于十分类数据集,针对本文模型的推理攻击成功率降低为 10%(推理攻击对 CNN 的攻击成功率约为 90%[2])。这说明,相比传统图像分类模型 CNN,本文方法加强了机器学习模型对推理攻击的防御能力。

2 预备知识

2.1 影子模型

Shokri 等首次提出了影子模型训练技术[3],借助影子模型,在推理攻击中生成目标模型原始训练数据的近似训练集。Ahmed 等利用影子模型训练技术,生成差分训练数据,进一步训练推理攻击模型[2]。

图 1 给出了 Ahmed 应用影子模型训练技术生成训练数据的方法。其中,Probe Set 为包含一组固定样本的探测集,Target Model 为推理攻击目标模型的近似模型,Shadow Model 为影子模型,每个影子模型与目标模型分别相差一次新样本的训练。将探测集输入目标模型和所有影子模型,得到多个差分向量结果 y,并将由每个影子模型产生的结果 y 分别与由目标模型产生的结果 $\hat{y}$ 相减得到差分向量 $\boldsymbol{\delta}$,如式(1)所示:

$$\boldsymbol{\delta}=y-\hat{y} \tag{1}$$

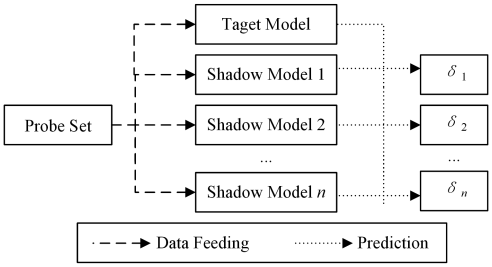


图 1 影子模型训练技术的训练数据生成方法

Fig. 1 Training data generation method of shadow model training technology

每个差分向量分别对应影子模型和目标模型的训练集所相差的训练样本,因此每个差分向量与其对应样本的标签将组成新的真实训练数据,可用于推理攻击模型训练。

2.2 ε-差分隐私

一个模型或查询函数的隐私保护能力常使用隐私预算 ε 来量化表示,ε 的值越小,模型的查询结果所暴露的数据集信息就越少,该模型的隐私性安全性也就越高。ε-差分隐私[4]的定义如下:

若两个只存在一条记录不相同的相邻数据集 D_1 和 D_2,以及一个机制 M 所产生的输出 R_1 和 R_2,满足式(2),则该机制 M 满足 ε-差分隐私。

$$\Pr(R_1\in S)\leqslant e^{\varepsilon}\times\Pr(R_2\in S) \tag{2}$$

其中,S 是所有可能的输出结果的子集。

3 相关工作

目前,机器学习的隐私保护主要通过差分隐私算法及其各种改进方案来实现。基于差分隐私算法的改进方案可分为 3 类:梯度级、函数级、标签级[1]。无论是哪种差分隐私算法,其本质都是通过各种方式在机器学习的过程中以不同的方式和策略添加噪声,以扰动神经网络对真实训练数据的记忆程度。

Abadi 等提出了 DP-SGD 算法[5],它通过计算神经网络权重参数对训练数据的依赖性,来控制梯度下降过程中的变化量以实现隐私保护,这种方式不利于复杂模型的收敛。Xie 等提出了 DP-GAN[6],通过在 Wasserstein 距离梯度上添加噪声来保护生成对抗网络的训练数据隐私,这种方式依赖于生成器生成高质量的训练数据点,因此不适用于更复杂的数据集。

Phan 等提出针对深层自动编码器的差分隐私机制[7]。在该机制中,他们为编码器添加额外的基于梯度下降算法的全局灵敏度计算层,以产生最佳扰动参数,随后使用反向传播算法对模型参数进行微调。这种方式引入了额外的、无关于拟合训练数据的神经层,对模型准确度的影响较大,在复杂数据集中的负面效果更加明显。作为改进方案,他们提出了一种新的差分隐私保护机制 ADLM[8],该机制通过向目标函数引进神经网络的参数,在训练过程中增大与输出弱相关的神经元上的噪声,动态地调整噪声分配。在 CIFAR-10 数据集上应用这种机制,在提高隐私安全性的同时,将模型准确度提升到 84.8%,相比 DP-SGD 算法的准确度提高了 14%,但该准确度仍不够理想。

Papernot 等提出了一种基于半监督知识迁移的深度学习训练策略[9]。该策略通过使用不相交数据集训练多个教师模型,随后以投票和添加噪声的方式聚合多个教师的预测结果,以进一步监督学生模型的训练,这种训练方式能够达到相对较高的模型准确度并拥有良好的隐私保护能力。但由于学生模型的训练数据依赖于由教师模型集合预测结果的聚合数据的质量,因此,对于复杂任务而言,需要大量的不相交数据训练多个具有极高准确度的教师模型,这是一个巨大的挑战。对于不同数据集,在聚合过程中如何设计噪声同样是一项复杂的任务。

总的来说,差分隐私算法应用于机器学习模型能有效地提高数据隐私安全性,但通过添加噪声的方式不利于模型拟合数据样本的特征规律,因为它以牺牲数据真实性为代价换取隐私安全,将不可避免地导致机器学习模型的分类能力下降。

4 本文方法

本文提出了一种基于特征映射的差分隐私保护机器学习方法，该方法采用多个预训练神经网络和一个主神经网络级联的结构进行机器学习训练。在训练阶段，当使用一个训练样本更新该模型时，每个预训练神经网络都将产生一个影子模型，每个预训练网络和各自的影子模型之间只存在与该训练样本相关的更新差异。然后，使用一组固定的探测集来探测每个预训练网络-影子模型对，得到的数量等于预训练神经网络数量的差分向量，将所有差分向量拼接组成一个高维差分向量，该高维差分向量将和原训练样本的标签组成一个新的训练样本。最后，使用该训练样本训练主神经网络。由于预训练网络使用包含原数据样本的数据集的不相交子集进行预训练，主神经网络使用由高维差分向量和原数据样本的标签进行训练，因此预训练神经网络和主网络拥有不同的结构。在分类预测阶段，模型使用相同的方式产生高维差分向量，并将高维差分向量输入主神经网络执行预测，得到分类结果。高维差分向量的特征传递过程在黑盒中隐藏，对用户不可见。

本文方法是借助预训练神经网络，将原始特征以差分向量的形式映射到高维向量空间上。映射后的同一分类的高维差分向量与原特征向量相比，分布更加密集，向量空间中每个分类簇内的向量之间拥有更短的距离，因此主神经网络能更好地拟合样本特征规律，使模型拥有更好的泛化能力。同时，由此产生的后验概率向量分布也将更密集，同一分类的后验概率向量间产生的差分向量数值极小，这意味着更少的样本数据隐私的泄露，因此该模型拥有更强的隐私保护能力。

本文方法的结果如图 2 所示，其中，Update Sample 为一个训练样本，Probe Set 为一组固定的样本集合，PNN 为预训练神经网络，Shadow PNN 为预训练神经网络接收一个训练样本更新所产生的影子模型，$\boldsymbol{\delta}$ 为差分向量，MNN 为主神经网络，Result 为主神经网络输出的后验概率向量。

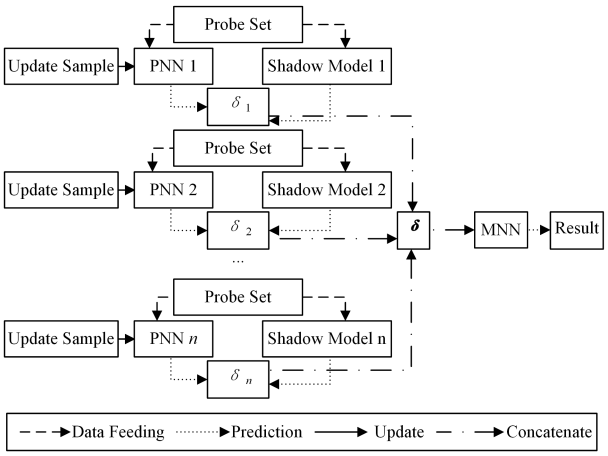


图 2 基于特征映射的神经网络结构

Fig. 2 Neural network structure based on feature mapping

在机器学习模型训练前，先将数据集划分成多个部分。首先，用 n 个不相交的、含一定数量样本的数据集子集分别对 n 个神经网络进行预训练，将第 i 个预训练神经网络记为 PNN i。这些神经网络的样本和训练轮数应该设置得较少，使它们学习到数据集特征的初步知识（这 n 个预训练神经网络都是训练不充分的神经网络，在未来的更新过程中，它们是固定的、不再被训练的，因为它们仅用于接收新样本输入，产生影子模型，并经过探测和拼接操作获得差分向量）。当模型接受到一个更新样本输入时，利用该更新样本平行地更新这 n 个预训练神经网络，得到 n 个影子模型，如图 2 中的 Shadow PNN$(1,2,\cdots,n)$。随后通过固定的探测集对 n 个预训练神经网络以及各自的影子模型进行探测得到 n 个差分向量。其中，探测集可以是原始数据集划分得到的一组小样本集合，也可以是通过一定策略生成的随机样本，本文实验中的探测集是根据输入维度随机生成的一组随机数值向量，同样取得了良好的效果。最后，将 n 个 $\boldsymbol{\delta}$ 向量进行拼接组成一个高维度向量，此时，该高维向量对应整个模型的最初接受的更新样本，因此该高维度向量和该更新样本的标签将组成一个新的数据样本，用于对主神经网络进行训练。

该训练过程的数学表达式如下：

$$y=M(\boldsymbol{\delta}) \tag{3}$$

$$\boldsymbol{\delta}=C(\delta_1,\delta_2,\cdots,\delta_n) \tag{4}$$

$$\boldsymbol{\delta}_i=P_i(D_{\text{probe}})-P_i'(D_{\text{probe}}) \tag{5}$$

其中，y 表示主神经网络 MNN 关于差分向量 $\boldsymbol{\delta}$ 的预测结果，M 是主神经网络 MNN 的函数表示，C 表示 n 个差分向量的拼接函数，D_{probe} 为探测集，第 i 个差分向量 $\boldsymbol{\delta}_i$ 表示第 i 个预训练神经网络 P_i 及其更新后的 P_i' 关于探测集的差分预测向量。

在式(3)中，输出向量 $\boldsymbol{y}$ 为模型接收输入样本最终产生的后验概率向量，它是由差分向量 $\boldsymbol{\delta}$ 输入主神经网络直接产生的。式(4)表示差分向量 $\boldsymbol{\delta}$ 由多个预训练神经网络和探测集 D_{probe} 探测得到的多个差分向量的拼接操作。式(5)表示探测集分别输入每个预训练神经网络和各自的影子模型得到输出结果差值产生差分向量的过程。

本文方法具有以下 3 个优势：

(1)隐私安全。现有的差分隐私算法通过添加噪声的方式，在提供隐私保护能力的同时会不可避免地导致模型分类能力下降。本文方法相比现有差分隐私算法在隐私保护问题上具有两个优势。首先，通过多个预训练神经网络，将包含样本信息的原始特征映射为更高维度的差分向量，以帮助主神经网络学习到更多的真实数据知识。同时，映射过程对原始样本的特征进行高维编码得到能够包含更多关于数据规律的特征，映射后的属于同一类别的差分向量在向量空间中的分布将更加密集，在推理过程中，该模型也将以同样的机制将接收到的输入映射到该密集区域中。因此，该模型拥有更好的泛化能力和数据隐私安全性。

(2)模型安全。Ahmed 的推理攻击和重构攻击以黑盒模型为攻击目标，依赖于一系列针对机器学习模型黑盒的攻击研究基础[10-14]，用于将目标黑盒模型转化为白盒模型。该前提条件使攻击者能够根据大量的输入以及黑盒模型的输出还原出数据集的近似分布。本文方法将机器学习黑盒模型从内部通过差分向量级联的方式分割成预训练神经网络和主神经网络，使前者的输出和后者的输入对外界隐藏。在攻击者对黑盒提供输入时无法获得预训练神经网络的输出，当他获得

黑盒输出结果时无法获得主神经网络的输入，因此他分别针对预训练神经网络和主神经网络都缺少还原数据集近似分布的条件，通过这种分割方式来保护整个机器学习黑盒模型的安全。同时，预训练神经网络在训练过程中是固定的、不被更新的，因此，从广义上而言，预训练神经网络被视为机器学习黑盒模型的超参数，为服务提供方私有，服务提供方能够对预训练神经网络采取额外的保护策略以提高模型的安全性。对于合法用户和恶意用户而言，整个模型是统一的，其功能和传统机器学习模型没有差别，用户不会感知到预训练神经网络的存在。

(3)可扩展。现实中的机器学习任务内容多种多样，数据集的复杂程度也各有不同。MNIST 在隐私保护研究中常作为基准数据集，被应用于模型的性能测试等任务中。由于 MNIST 数据集中样本的简单性，本文实验仅使用一个预训练神经网络，训练后的机器学习模型就能达到很高的分类准确度，但现实中其他任务所处理数据的维度可能远远超过 MNIST 样本的维度，一个预训练神经网络不能使模型达到预期的分类准确度，因此该模型可以利用不相交的多个数据集子集，训练数量可调的预训练神经网络，以扩展差分向量的维度，使差分向量能包含更多原始数据特征，从而提高模型的分类准确度。因此，这种多预训练神经网络的结构设计，使本文方法对不同的分类任务具有可扩展性。

5 实验及结果分析

5.1 实验环境

本文使用隐私保护研究时，常用基准数据集 MNIST 和 CIFAR-10，设置了两组实验验证本文模型的图像分类能力、隐私保护能力和对推理攻击的防御能力。

第一组实验使用 MNIST 数据集。首先，用预训练神经网络训练样本数量，用主神经网络训练样本数量，用推理攻击模型训练样本数量，将各网络训练轮数作为参数，用本文方法训练多个模型。然后，使用 Ahmed 提出的单样本标签推理攻击方式攻击该模型并计算攻击成功率，以推理攻击成功率的数值量化该模型对推理攻击的防御能力，攻击成功率越高，该模型对推理攻击的防御能力就越弱，攻击成功率越低，该模型对推理攻击的防御能力就越强，计算该模型的隐私预算 ε 和分类准确度以量化该模型的隐私保护能力和分类能力。实验中的探测集使用一组固定的 100 个随机生成的 784 维向量集合。

图 3 给出了 Ahmed 的攻击方法的攻击模型[2]，其中 ML Black-Box 表示被攻击的机器学习黑盒模型。将 ML Black-Box 作为目标模型，并使用影子模型训练技术完成攻击神经网络的训练。当 ML Black-Box 接收新样本时会进行更新，生成该模型的更新版本，如图 3 中的 Updated ML Black-Box。随后使用探测集探测两个模型的差异从而产生差分向量 $\boldsymbol{\delta}$，最后将 $\boldsymbol{\delta}$ 向量作为输入传入攻击神经网络从而产生预测结果。若该预测结果和该新样本的标签对应则攻击成功，否则攻击失败。用相同的方式使用一组数据集进行测试并统计成功次数，将结果记为攻击成功率(attack success rate)。攻击成功率越高，该模型对推理攻击的防御能力就越弱，攻击成功率越低，该模型对推理攻击的防御能力就越强。

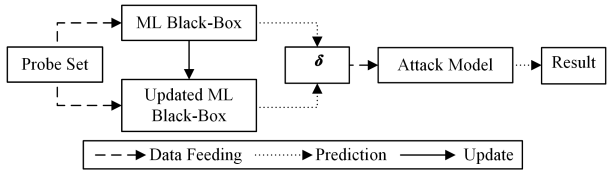


图 3 攻击模型

Fig. 3 Attack model

第一组实验使用一个预训练神经网络，其结构使用包含两个核为 5 的二维卷积层、两个核为 2 的池化层以及两个全连层的卷积神经网络，激活函数使用 ReLU 函数。当模型训练完成后，将其作为机器学习黑盒，并使用 Ahmed 的攻击方法对其进行攻击，比较该模型的分类准确度，以及不同拟合程度的预训练神经网络下的推理攻击攻击成功率。

第二组实验使用 CIFAR-10 数据集。CIFAR-10 数据集中，每张图片共含 3＃072 个特征，简单使用一个预训练神经网络并不能达到很好的效果，因此通过添加预训练神经网络的数量，进一步扩展主神经网络的输入向量维度以获得更好的训练效果。在实验中取 4 个预训练神经网络，将映射后的特征扩展到 4000 维(若使用 1 个预训练神经网络，探测得到的差分向量为 1000 维)。该组实验以相同的方式测试本文方法训练的模型的隐私保护能力、分类能力及对推理攻击的防御能力。实验中的探测集使用一组固定的 100 个随机生成的 3072 维向量集合。

第二组实验中，预训练神经网络包含两个核为 5 的二维卷积层、两个核为 2 的池化层以及 3 个全连接层的卷积神经网络。主神经网络包含 7 个全连接层，将 4000 维差分向量降至 10 维并输出，激活函数使用 ReLU 函数。每个预训练神经网络使用 5000 个不相交的训练样本训练 10 个 epoch，主神经网络使用 10000 个训练样本训练 50 个 epoch。最后，同样使用图 3 所示的攻击模型和推理攻击方式对模型进行攻击，根据统计获得模型的分类准确度和推理攻击成功率

5.2 实验结果分析

5.2.1 模型参数对分类准确度及对推理攻击防御能力的影响

如表 1 所列，取 10000，20000，30000 个样本分别对预训练神经网络 PNN 训练 10，25，50 个 epoch，得到模型的分类准确度以及针对该模型的推理攻击的成功率。

表 1 MNIST 数据集实验中，不同参数下应用本文方法训练模型的分类准确度和攻击准确度对比

Table 1 Comparation about the classification accuracy and attack accuracy of the training model under different parameters in MNIST dataset

PNN Size/个	PNN Epoch/轮	Model Accuracy/%	Attack Success Rate/%
10000	10	99.51	11.60
20000	10	90.67	23.80
30000	10	92.11	24.17
10000	25	86.17	27.11
20000	25	59.82	44.53
30000	25	54.89	45.34
10000	50	55.42	57.21
20000	50	26.83	74.44
30000	50	22.93	61.50

观察模型准确度和针对该模型的推理攻击成功率的数据

可以发现，两者之间存在明显的负相关关系，这验证了推理攻击的基础来自机器学习模型的过拟合，并且随着模型泛化能力的降低，对推理攻击的防御能力逐步变弱。当 PNN Size 为 10 000、epoch 为 10 时，整个模型的分类准确度达到了 99.51%，模型的分类性能达到最优，对推理攻击的防御能力也最好。随着预训练样本数量的增加或预训练神经网络训练 epoch 的增加，整个模型的分类准确度下降，过拟合程度变高，针对该模型的推理攻击成功率变高，该模型对推理攻击的防御能力变弱。

当预训练神经网络的训练样本数量 PNN Size 为 10 000、训练轮数 PNN Epoch 为 10 时，针对该模型的推理攻击的成功率降至 11.6%。由于 MNIST 为 10 分类数据集，该攻击效果等同于随机猜测的攻击效果，这意味着在该参数条件下训练的基于知识差异的隐私保护机器学习模型将使针对该模型的推理攻击失效（由于 MNIST 为 10 分类任务，因此成功率 11.6%的攻击效果近似于随机猜测）。

表 2 列出了在 MNIST 数据集上，CNN 模型与本文模型在不同训练集大小和训练轮数上的模型分类准确度及针对这两种模型的推理攻击成功率。表 3 列出了在 CIFAR 数据集上，CNN 模型与本文模型在不同训练集大小和训练轮数上的模型分类准确度及针对这两种模型的推理攻击成功率对比。模型的分类准确度越高，该模型的泛化能力越高，预测性能就越好。针对该模型的推理攻击的成功率越高，该模型对推理攻击的防御能力就越差。

表 2 MNIST 数据集实验中，在不同参数下本文方法和 CNN 模型的分类准确度和针对该模型的推理攻击准确度的对比

Table 2 Comparation about the classification accuracy and attack accuracy between our method and CNN under different parameters in MNIST dataset

Dataset Size/个	Epoch/轮	Model Accuracy (本文方法)/%	Attack Success Rate (本文方法)/%	Model Accuracy (CNN)/%	Attack Success Rate(CNN)/%
10000	10	99.32	9.80	97.45	99.97
10000	25	99.56	10.35	98.07	90.32
10000	50	99.74	11.48	92.43	64.41
15000	10	99.57	11.45	98.23	81.64
15000	25	99.84	11.66	98.52	69.07
15000	50	99.92	9.75	95.78	52.01
20000	10	99.82	10.31	98.64	79.34
20000	25	99.95	10.26	98.67	60.34
20000	50	99.99	11.46	96.13	50.49

表 3 CIFAR 数据集实验中，在不同参数下本文方法和 CNN 模型的分类准确度和针对该模型的推理攻击准确度对比

Table 3 Comparation about the classification accuracy and attack accuracy between our method and CNN under different parameters in the CIFAR dataset

Dataset Size/个	Epoch/轮	Model Accuracy (本文方法)/%	Attack Success Rate (本文方法)/%	Model Accuracy (CNN)/%	Attack Success Rate(CNN)/%
10000	10	96.46	11.21	56.38	94.70
10000	25	97.79	9.76	68.14	96.82
10000	50	98.83	11.63	70.54	97.13
15000	10	97.75	11.61	60.42	95.41
15000	25	97.88	9.78	67.37	97.52
15000	50	98.98	11.03	71.92	98.13
20000	10	97.34	9.08	63.88	96.88
20000	25	99.30	10.10	70.42	97.17
20000	50	99.32	9.51	73.56	97.56

表 2 中，Dataset Size 参数对于本文方法而言表示用于 MNN 训练的样本数量，不包含 PNN 的预训练样本；对于 CNN 而言表示训练整个模型的训练样本数量。epoch 参数对于本文方法而言表示 MNN 的训练轮数，对于 CNN 而言表示整个模型的训练轮数。

如表 2 所列，对于 MNIST 和 CIFAR-10 数据集，本文方法与 CNN 相比，在相同训练集大小（不包括预训练数据）和训练轮数时，能保持更高的分类准确度，并有效提高对推理攻击的防御能力，使推理攻击的准确度降低到 12%以下。在 MNIST 数据集中，针对 CNN 模型的推理攻击，当 CNN 模型的训练轮数为 50 轮时，攻击准确度将发生显著降低，但此时 CNN 模型发生过拟合，其泛化能力变弱，分类准确度也降低。

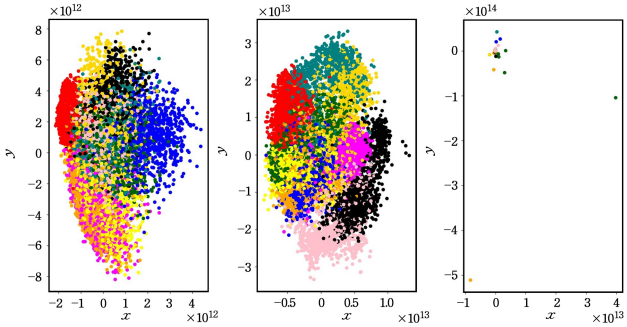
5.2.2 特征映射效果分析

对 MNIST 实验中的各个中间向量进行主成分分析，使其能在二维坐标系上可视化，如图 4、图 5 所示。

图 4 给出了预训练样本为 10000、预训练轮数为 10、主神经网络训练样本为 10000、主神经网络训练轮数为 50、模型分类准确度为 99.74%、针对该模型的推理攻击成功率为 11.48%的实验组的各向量数据分布，该实验组具有较高的对推理攻击的防御能力和隐私保护能力。图 5 给出了预训练样本为 10000、预训练轮数为 50、主神经网络训练样本为10000、主神经网络训练轮数为 50、模型分类准确度为 64.41%、针对该模型的推理攻击成功率为 57%的实验组的各向量数据分布，该实验组具有较弱的对推理攻击的防御能力和隐私保护能力。

其中，样本点的颜色表示该样本点的分类，共有 10 类。图 4(a)、图 5(a)表示原始数据集在向量空间上的分布，图 4(b)、图 5(b)表示训练样本的原始特征向量借助预训练神经网络生成的高维差分向量在向量空间中的分布，图 4(c)、图 5(c)表示由攻击模型探测本文方法训练所得模型得到的后验

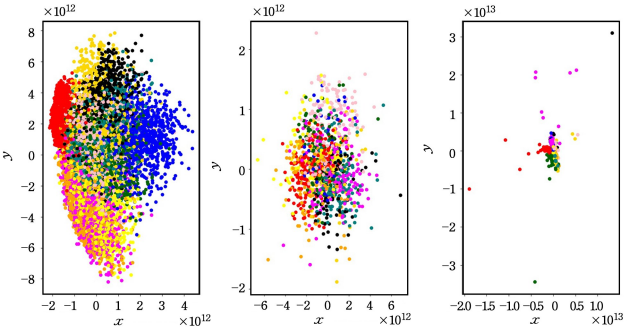
差分向量在向量空间上的分布，由于该差分向量包含被攻击模型的更新样本的信息，因此该差分向量在向量空间上的分布将反映被攻击模型的隐私泄露程度，该差分向量越远离零点，向量内各维度的数值越高，其包含的更新样本信息越多，该向量所泄露的隐私信息就越多。因此，在图 4(c)、图 5(c)中，远离零点的差分向量样本点数量越多，该模型的隐私保护能力就越差。



(a) Distribution of Dataset (b) Distribution of Delta (Input of MNN) (c) Distribution of Delta (Input of Attack Model)

图 4 高隐私保护能力实验组的各向量数据分布(电子版为彩色)

Fig. 4 Distributions of vectors in the experimental group with high privacy protection ability



(a) Distribution of Dataset (b) Distribution of Delta (Input of MNN) (c) Distribution of Delta (Input of Attack Model)

图 5 低隐私保护能力实验组的各向量数据分布(电子版为彩色)

Fig. 5 Distributions of vectors in the experimental group with low privacy protection ability

观察图 4 中各向量在向量空间中的分布可以发现，原始训练集在向量空间上大致可分为 10 个簇，但簇与簇之间存在较多的重叠。在高隐私保护能力实验组中，借助预训练神经网络生成的高维差分向量在向量空间上的同一分类的高维差分向量分布更加密集，簇与簇之间的界限更加明显。新样本经过同样的方式映射到分布密集的类簇中，使用映射后的高维差分向量训练得到的机器学习模型将拥有更好的泛化能力。在高隐私保护能力实验组中，基于推理攻击的攻击模型针对该模型探测得到的差分向量集中于零点邻域，这意味着该差分向量由于数值过小，包含极少的更新样本的隐私泄露信息，这直接导致了推理攻击的失效。因此该实验组中，该分类模型具有更高的分类准确度和更高的对推理攻击的防御能力，即更高的隐私保护能力。

图 5 给出了低隐私保护能力实验组的各向量在向量空间中的分布。它与高隐私保护能力实验组相比，预训练神经网络的训练轮数由 10 轮增加到 50 轮，这导致预训练神经网络的拟合程度过高，由原训练样本到高维差分向量的特征映射效果下降，如图 5(b)所示。这导致攻击模型对该模型探测得到的差分向量偏离零点，这些差分向量携带更多的更新样本的信息，因此针对该模型的推理攻击成功率上升为 57%，表明该实验组中的分类模型具有更低的对推理攻击的防御能力和隐私保护能力。

上述分析验证了本文方法能通过特征映射的方式提高模型的隐私保护能力。同时，通过控制预训练神经网络的拟合程度能进一步调整该模型的隐私保护能力。

5.2.3 多种算法的性能对比

表 4 列出了多种隐私保护算法在 MNIST 和 CIFAR-10 数据集上的分类准确度和隐私预算对比。其中，隐私预算是模型隐私保护能力的量化表示，ε 越小，模型的隐私保护能力就越好，用分类准确度衡量模型的预测能力，分类准确度越高，模型的分类预测能力就越好。在 MNIST 数据集的实验中，本文方法训练的 ε 为 0.01 的 ε-差分隐私模型，准确度达到了 99%。在 CIFAR-10 数据集的实验中，本文方法训练的 ε 为 0.11 的 ε-差分隐私模型，准确度达到了 96%。该实验结果表明，本文方法优于 DP-SGD 等多种常用机器学习隐私保护算法。

表 4 多种隐私保护算法及本文方法应用于 MNIST 和 CIFAR-10 数据集时的分类准确度和隐私保护能力对比

Table 4 Comparation about the classification accuracy and capabilities of privacy protection between various algorithms and our method in MNIST and CIFAR-10

算法	算法简要描述	隐私预算 ε (MNIST)	分类准确度 Accuracy/% (MNIST)	隐私预算 ε (CIFAR-10)	分类准确度 Accuracy/% (CIFAR-10)
本文方法	不添加噪声，而是借助预训练神经网络将原始特征映射到差分向量，并用于训练主神经网络	0.01	99	0.11	96
pCDBN[15]	对基于能量的目标函数添加噪声	0.2	58	—	—
RBADN[16]	根据不同神经元相关性和模型的输出添加噪声	0.2	75	2	68
ADLM[8]	根据特征贡献度向输入添加噪声	0.25	91	0.5	84
GDP[17]	基于统计方法优化噪声的分布，使输出添加噪声后更接近而难以区分	0.83	95	—	—
PATE[9]	基于知识聚合、知识迁移和噪声机制训练教师-学生模型	1.9	98	—	—
DP-SGD[5]	基于训练数据向梯度添加噪声	8	97	8	73

结束语 本文提出了一种基于特征映射的差分隐私保护机器学习方法。该方法提供隐私保护能力的方式与现有隐私保护算法不同，其在训练和预测过程中不添加噪声，而是结合预训练神经网络和影子模型训练技术，以差分向量的形式映射到高维向量空间，缩短样本点在向量空间中的距离，进一步减少模型更新所泄露的差分隐私信息，从而达到隐私保护的目的。同时，本文方法将预训练神经网络和主神经网络级联相接，当作为黑盒对外提供服务时，这种级联方式能为模型提供结构安全保障，加大了攻击者构建近似模型的难度。在MNIST和CIFAR-10数据集的应用中，本文方法训练ε分别为0.01和0.11的ε-差分隐私模型，其分类准确度提高到了99%和96%，相比DP-SGD等多种差分隐私保护算法，该模型能在更低的隐私保护预算下达到更高的分类准确度。同时，在MNIST和CIFAR-10数据集上，应用了本文方法的模型对基于差分的推理攻击的防御能力相比传统图像分类模型CNN具有较大幅度的提升。

在本文方法中，预训练神经网络需要大量额外的数据样本。另外，本文实验仅将该方法应用于计算机视觉中的十分类任务。因此，如何减少该方法对数据样本的需求量，以及对于该方法是否适用于其他计算机视觉任务和语音、文本处理等任务，将是未来进一步的研究方向。

参 考 文 献

[1] HA T, DANG T K, DANG T T, et al. Differential Privacy in Deep Learning: An Overview[C]// 2019 International Conference on Advanced Computing and Applications (ACOMP). Piscataway, NJ, USA: IEEE, 2019: 97-102.

[2] AHMED S, APRATIM B, MICHEAL B, et al. Updates-Leak: Data Set Inference and Reconstruction Attacks in Online Learning[C]// 29th USENIX Security Symposium. Online: USENIX Association, 2019: 1291-1308.

[3] SHOKRI R, STROATI M, SONG C Z, et al. Membership Inference Attacks Against Machine Learning Models[C]// 2017 38th IEEE Symposium on Security and Privacy (SP). Los Alamitos, CA, USA: IEEE Computer Society, 2017: 3-18.

[4] DWORK C, KENTHAPADI K, MCSHERRY F, et al. Our data, ourselves: privacy via distributed noise generation[C]// 24th Annual International Conference on the Theory and Applications of Cryptographic Techniques Advances in Cryptology (EUROCRYPT 2006). Berlin, Germany: IEEE Computer Society, 2006: 486-503.

[5] ABADI M, MCMAHANH B, CHU A, et al. Deep learning with differential privacy[C]// Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security (CCS 2016). Vienna, Austria: Association for Computing Machinery, 2016: 308-318.

[6] XIE L Y, LIN K X, WANG S, et al. Differentially Private Generative Adversarial Network[J/OL]. http://arxiv.org/abs/1802.06739, 2020-5-13.

[7] PHAN N, WANG Y, WU X, et al. Differential privacy preservation for deep auto-encoders: An application of human behavior prediction[C]// 30th AAAI Conference on Artificial Intelligence (AAAI 2016). Phoenix, AZ, United states: AAAI press, 2016: 1309-1316.

[8] PHAN N, WU X, HU H, et al. Adaptive Laplace mechanism: differential privacy preservation in deep learning[C]// 2017 IEEE International Conference on Data Mining (ICDM). Los Alamitos, CA, USA: IEEE Computer Society, 2017: 385-394.

[9] PAPERNOT N, GOODFELLOW I, ABADI M, et al. Semi-supervised knowledge transfer for deep learning from private training data[C]// 5th International Conference on Learning Representations (ICLR 2017). Conference Track Proceedings. Toulon, France: ICLR, 2017: 1024-1040.

[10] GANJU K, WANG Q, YANG W, et al. Property inference attacks on fully connected neural networks using permutation invariant representations[C]// Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security (CCS 2018). United States: Association for Computing Machinery, 2018: 619-633.

[11] JOON O S, BERNT S, MARIO F. Towards Reverse-Engineering Black-Box Neural Networks[J]. Springer Verlag, 2017, 11700(2017): 121-144.

[12] SALEM A, YANG Z, HUMBERT M, et al. ML-Leaks: Model and Data Independent Membership Inference Attacks and Defenses on Machine Learning Models[J/OL]. http://arxiv.org/abs/1806.01246, 2018-12-14.

[13] SHOKRI R, STRONATI M, SONG C, et al. Membership Inference Attacks Against Machine Learning Models[C]// 2017 38th IEEE Symposium on Security and Privacy (SP). Los Alamitos, CA, USA: IEEE Computer Society, 2017: 3-18.

[14] WANG B, GONG N. Stealing Hyperparameters in Machine Learning[C]// 2018 IEEE Symposium on Security and Privacy (SP). Los Alamitos, CA, USA: IEEE Computer Society, 2018: 36-52.

[15] PHAN N, WU X, DOU D. Preserving differential privacy in convolutional deep belief networks[J]. MACH LEARN, 2017, 106: 1681-1704.

[16] GONG M, PAN K, XIE Y, et al. Preserving differential privacy in deep neural networks with relevance-based adaptive noise imposition[J]. Neural Networks, 2020, 125: 131-141.

[17] DONG J S, ROTH A, SU W J, et al. Gaussian Differential Privacy[J/OL]. http://arxiv.org/abs/1905.02383, 2019-10-08.

CHEN Tian-rong, born in 1996, postgraduate. His main research interests include digital image processing and privacy protection.

LING Jie, born in 1964, Ph.D, professor. His main research interests include information security technology and intelligent video processing technology.