

引用格式:侯小军,李泽华,李泽堃.支持层级相关性传播的差分隐私分类算法研究[J].微电子学与计算机,2021,38(5):48-53. [HOU X J, LI Z H, LI Z K. Research on differentially private classification algorithm based on layer-wise relevance propagation[J]. Microelectronics & Computer, 2021, 38(5): 48-53.]

支持层级相关性传播的差分隐私分类算法研究

侯小军¹, 李泽华², 李泽堃²

(1 西安理工大学 信息化管理处, 陕西 西安 710048; 2 陕西师范大学 计算机科学学院, 陕西 西安 710119)

摘 要: 为了防止攻击者在深度学习图像分类过程中还原训练集数据并保护输入图像数据, 提出一种基于层级相关性传播的差分隐私分类算法. 该算法首先采用层级相关性传播模型量化图像的特征相关性, 然后利用相关性自适应地向损失函数添加噪声并利用 Adam 算法进行模型优化, 最后依据相关性分配隐私预算并构造差分隐私变换层以扰动输入数据. 实验结果表明, 该算法在实现隐私保护的同时, 能够保证较高的分类准确率.

关键词: 层级相关性; 深度学习; 差分隐私; 噪声; 分类

中图分类号: TP309

文献标识码: A

文章编号: 1000-7180(2021)05-0048-06

Research on differentially private classification algorithm based on layer-wise relevance propagation

HOU Xiao-jun¹, LI Ze-hua², LI Ze-kun²

(1 Xi'an Technological University, Information management department, Xi'an 710048, China;

2 School of Computer Science, Shaanxi Normal University, Xi'an 710048, China)

Abstract: To prevent attackers from restoring the training dataset and to protect the input image during the application process of deep learning image classification model, a differential privacy classification algorithm based on LRP (Layer-wise Relevance Propagation) is proposed in the paper. The relevance between image features is firstly quantified according to LRP in the proposed algorithm, then adaptive noise is added to the loss function based on feature relevance and Adam mechanism is used for model optimization. Finally, a differential private transform layer is constructed to perturb the input image and privacy budget is assigned according to feature relevance during model application stage. Experimental results show that the proposed algorithm achieves high classification accuracy in condition of privacy preservation.

Key words: layer-wise relevance propagation (LRP); deep learning; differential privacy; classification; privacy budget

1 引言

深度学习因其端到端的数据处理方式、强大的表征提取能力,有效缓解了海量图像数据场景下传统机器学习方法面临的处理过程复杂、识别效率低下等问题.深度学习在训练和应用过程中均存在较

为严峻的隐私泄露风险.应用服务商为了保证服务质量,很少对收集的用户数据进行隐私保护处理,直接对原始数据进行挖掘分析容易造成隐私泄露;用户在使用服务产生个人数据并提交的时候,通常也没有意识到可能存在的隐私泄露风险.本文针对图像分类问题,将深度学习中的卷积神经网络和差分

收稿日期: 2020-08-08; 修回日期: 2020-09-10

基金项目: 西安市科技计划项目(2019216914GXRC005CG006-GXYD5.2); 中央高校基本业务费项目(GK201906009); 赛尔网络下一代互联网技术创新项目(NGII20190704)

隐私模型^[1-3]、层级相关性传播模型相结合^[4],构建基于图像相关性的隐私保护分类模型并设计分类算法,在实现隐私保护的同时能够保证较高的分类准确率。

深度学习的隐私风险可产生在训练阶段和预测阶段^[5],预测阶段的隐私风险不考虑训练过程中因攻击者恶意参与而导致的数据泄露问题,指训练完成的深度学习模型在公开模型结构及参数,提供接口供外部使用时,攻击者能够通过模型反演攻击^[6]等手段间接获取训练数据集,从而导致隐私泄露。对深度学习的差分隐私加噪模型包括在迭代步骤中梯度裁剪和添加高斯噪声^[7]、采用函数机制扰动其损失函数的近似多项式系数^[8]、为对抗网络的判别器梯度计算中添加高斯噪声^[9-10]等实现隐私保护,综合现状可知当前基于差分隐私的深度学习研究偏重于解决预测阶段的隐私风险,但是当前研究对自适应的隐私预算分配和数据的输入扰动考虑较少。

本文综合考虑深度学习使用场景,提出 DP-LRPCA (Differentially Private LRP Classification Algorithm) 算法,该算法利用层级相关性传播模型量化输入特征与分类结果之间的相关性,并将预测阶段细分为模型生成阶段和模型应用阶段,在模型生成阶段利用相关性自适应计算迭代过程所需的隐私预算用于添加噪声,在模型应用阶段利用相关性向输入特征自适应分配隐私预算,解决图像训练数据集和应用数据的隐私保护问题,同时自适应方式实现不同程度的扰动以提高模型分类准确率,保证隐私保护和精准分类之间的平衡。

2 相关定义

定义 1(差分隐私 -Differential Privacy) 设有随机函数 M ,其取值范围为 P_M 。对于任意两个邻近数据集 D 和 D' ,在 P_M 的任意一个子集 S_M 中,若函数 M 满足以下不等式,则称 M 满足 ϵ - 差分隐私。

$$Pr[M(D) \in S_M] \leq \exp(\epsilon) \times Pr[M(D') \in S_M] \quad (1)$$

式中, $Pr[x]$ 表示隐私数据 x 风险披露的概率,参数 ϵ 称为隐私预算,用来衡量隐私保护的强度, ϵ 越小,隐私保护程度越高。

定义 2 (层级相关性传播模型-Layer-wise Relevance Propagation, LRP) 层级相关性传播模型是一种通过在神经网络中运行反向传递机制来识别重要图像特征的方法,可以看作是对预测函数的深度泰勒分解。在反向传递的过程中,不断利用上一层神

经元所计算的相关性进行迭代,然后通过权重等信息重新计算本层相关性,最终将相关性传递至输入层,并获取不同输入特征的相关性。

3 算法设计

3.1 架构设计及设计思路

DP-LRPCA 算法采用集中式学习架构,不考虑攻击者非法访问数据集及恶意参与者造成的隐私泄露问题,重点解决因模型结构及参数公开、提供接口供外部调用场景下的隐私泄露问题。DP-LRPCA 算法利用卷积神经网络模型用于图像分类,其中卷积神经网络模型结构即输入图像经过三组卷积与池化之后,经过全连接映射到 10 维的输出,并对其进行 Softmax 处理作为最终输出。

DP-LRPCA 算法综合考虑模型生成和应用阶段的隐私保护,首先利用预训练好的神经网络进行反向传播从而量化输入特征与分类结果之间的相关性。其次,在模型生成阶段,除了利用上述卷积神经网络学习训练数据集的分布特点之外,同时利用相关性计算结果生成每一次迭代过程的隐私预算,用于扰动梯度计算过程以实现对训练数据集的隐私保护。最后,在模型应用阶段,利用相关性计算结果分配隐私预算,并对图像每一个维度的输入特征进行扰动以实现隐私保护图像分类。

相关性计算的主要思路如下:读取批量图像数据并进行特征维度上的相关性计算,记为 R_i ,表示第 i 维输入特征与分类结果的相关性,并将当前迭代中所有维度的特征相关性累加求均值,记为 $R_{(t)}$,表示第 t 次迭代的数据分布,迭代多次之后,将特征维度上的相关性均值记为 $\bar{R}_i$,将所有迭代的数据分布均值记为 $\bar{R}$ 。

针对模型生成及应用过程的隐私保护,DP-LRPCA 算法分别在模型生成阶段和应用阶段利用相关性实现对训练数据集和应用数据集的隐私保护。在生成阶段,利用相关性计算隐私预算并扰动梯度计算过程;在应用阶段,利用相关性分配隐私预算扰动图像的输入特征。

3.2 模型生成阶段的隐私保护

模型生成阶段中的隐私保护主要通过迭代步骤中的梯度计算过程添加自适应的拉普拉斯噪声来实现,得到的扰动梯度用 g'_i 表示,如公式(2)所示。其中 L 表示迭代过程批量样本的数量, x_m 表示第 m 个训练数据, Δ_F 表示梯度求和函数的全局敏感度, ϵ_i

表示第 t 次迭代过程的隐私预算, $\text{Lap}(\Delta_F/\epsilon_t)$ 表示位置参数为 0, 尺度参数为 Δ_F/ϵ_t 的拉普拉斯噪声. 接着利用扰动后的梯度 g'_t , 采用 Adam 算法分别计算一阶矩、二阶矩的值并进行参数更新.

$$g'_t \leftarrow \frac{1}{L} (\sum_m g_t(x_m) + \text{Lap}(\Delta_F/\epsilon_t)) \quad (2)$$

为了方便确定敏感度, 本文遵从文献[8]的思路, 对每个样本的梯度进行裁剪, 并且设置裁剪阈值为 C , 通过分析可知, Δ_F 取值为 $2C$.

ϵ_t 是第 t 次迭代过程的隐私预算, 由相关性自适应计算得到, 其计算过程如下:

在卷积神经网络生成过程第 t 次迭代中, 基于层级相关性传播可计算得到每一维度的相关性 R'_i , 并进而得到当前迭代的数据分布值 $R'_{(t)}$. 利用下列公式(3) 计算得到当前迭代过程的数据分布占整个模型生成过程的比例.

$$P_t = \frac{R'_{(t)}}{\text{iter_num} * \bar{R}} \quad (3)$$

式中, iter_num 表示训练的最大迭代次数, $\bar{R}$ 是利用预训练好的神经网络计算得到的数据分布均值.

然后利用比例 P_t 获取当前迭代过程的自适应隐私预算 ϵ_t , 通过公式(4) 计算每次迭代所需要的隐私预算 ϵ_t :

$$\epsilon_t = \ln\left(\frac{1+P_t}{1-P_t}\right) \quad (4)$$

DP-LRPCA 算法不需要对隐私预算进行独立分配, 而是根据平均相关性的取值来量化整个模型, 并在模型生成过程中, 自适应地调整隐私预算的值. 这样能够提高隐私保护模型的分类型准确率. 从下文图 1 可以看出, 隐私预算与数据分布比例 P 的取值成正比, 当 P 的取值 $0 \sim 1$ 范围内, 隐私预算的取值为 $0 \sim 2.3$ 之间.

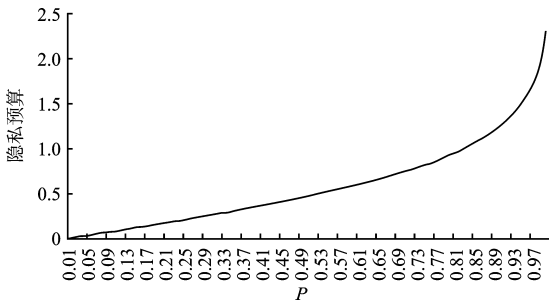


图 1 隐私预算与数据分布比例 P 的关系图

通过对每一次迭代过程中隐私预算的统计, 最终得出模型生成过程的整体隐私预算取值 ϵ_1 , 计算公式由(3)(4) 可得

$$\epsilon_1 = \sum_{t=1}^{\text{iter_num}} \ln\left(\frac{1+R'_t/\text{iter_num} * \bar{R}}{1-R'_t/\text{iter_num} * \bar{R}}\right) \quad (5)$$

需要注意的是, 在计算模型生成过程的隐私预算时, 隐私预算的取值会随着迭代过程中 $R'_{(t)}$ 取值的变换而不断改变. 因此, 对于隐私预算在生成模型部分的取值, 需要提前进行简单的预测, 即计算出的自适应隐私预算的总和需要小于等于生成阶段整体的隐私预算, 同时保证迭代过程累计数据分布比例之和不超过 1. 如果在迭代过程中超过提前分配好的隐私预算或者累计比例之和超过 1, 将不能在剩余的迭代过程中实施隐私保护.

3.3 模型应用阶段的隐私保护

在应用分类模型对应用数据集进行分类之前, 需要对图像数据的输入特征进行隐私保护, 构造差分隐私变化模块对每一个输入特征添加 Laplace 噪声来实现差分隐私. 具体来说, 对样本 x_m 的第 i 维输入特征 $N_i(x_m)$ 而言, 采用下列公式对其添加噪声:

$$N'_i(x_m) = N_i(x_m) + \frac{1}{|L|} \text{Lap}(\Delta_h/\epsilon_i) \quad (6)$$

式中, $|L|$ 表示应用过程中输入的样本数, ϵ_i 表示分配到第 i 维输入特征的隐私预算, ϵ_i 的取值公式为:

$$\epsilon_i = \epsilon_2 \times \frac{\bar{R}_i}{\sum_{i=1}^n \bar{R}_i} \quad (7)$$

在式(7) 中, $\bar{R}_i$ 表示第 i 维特征的相关性均值, n 表示样本维度, $\sum_{i=1}^n \bar{R}_i$ 表示对所有维度的相关性均值累加求和, ϵ_2 是模型应用阶段的整体隐私保护预算. 简单起见, 本文取 $\epsilon_2 = \epsilon_1$. 这里, 我们借鉴文献[11] 的函数敏感度公式来计算 Δ_h , 如下式所示:

$$\begin{aligned} \Delta_h &= \sum_{i=1}^n \left\| \sum_{x_m \in L} N_i(x_m) - \sum_{x'_m \in L'} N_i(x'_m) \right\|_1 \\ &\leq 2\max \sum_{i=1}^n ||x_{mi}||_1 \end{aligned} \quad (8)$$

式中, L 和 L' 是一对邻近数据集, x_m 和 x'_m 表示邻近数据集内唯一不同的记录, 则敏感度在于两条记录所有维度特征的最大差异程度, 此外, 算法对数据 x 进行归一化处理, 值被约束到 $[-1, 1]$, 因此敏感度为 $2\max \sum_{i=1}^n ||x_{mi}||_1$.

从公式(7)分析可知, 相关性越小, 其分配的隐私预算越小, 添加的噪声也就越大, 反之亦然. 上述对输入特征的扰动过程表明对图像中相关性较大的重要特征扰动程度小, 对相关性较小的输入特征扰动程度大, 因此对数据本身扰动程度小, 使得模型提

供隐私保护的同时,能够尽可能高地保证分类准确率。

3.4 DP-LRPCA 算法实现流程

DP-LRPCA 算法以卷积神经网络作为基础网络结构,分别在模型生成过程和模型应用过程添加拉普拉斯噪声进行扰动,以实现隐私保护。DP-LRPCA 算法整体上分为相关性计算、模型生成和应用分类三个阶段,首先确定预训练好的神经网络,可以是开源稳定模型的结构及参数,也可以是通过自主训练得到的稳定模型,利用预训练好的神经网络计算训练数据集的每个维度特征的平均相关性和整体数据分布的均值;接着将训练数据集作为输入数据集,采用层级相关性传播计算每次迭代过程的数据分布值及其比例,并按照比例自适应生成隐私预算并在梯度计算过程中加噪,经过多次迭代训练,生成稳定的网络模型;在应用分类阶段,输入应用数据集,利用相关性结果分配每一维的隐私预算,并按照特征维度对应用数据集加噪,最后利用加噪后的应用数据集进行图像分类。

DP-LRPCA 算法的总输入为函数敏感度 Δ_F 和 Δ_h ,训练数据集 D_1 ,应用数据集 D_2 ;总输出为通过隐私保护模型计算的分类准确率及分类结果;算法分为三个阶段,第一阶段进行相关性的计算,根据输入训练数据集 D_1 ,利用层级相关性传播模型进行反向传播,输出特征相关性均值 $\bar{R}_i$ 和数据分布均值 $\bar{R}$;第二阶段进行模型训练,达到最大迭代次数即终止,可细分为四个子步骤,该步骤结束后输出稳定的隐私保护图像分类模型:

(1) 从训练数据集 D_1 中随机抽取批量训练数据;

(2) 利用层级相关性传播计算迭代过程的数据分布 R'_t ;

(3) 利用公式(4)计算需要的自适应隐私预算的值 $\epsilon_t, \epsilon_1 = \epsilon_1 + \epsilon_t$;

(4) 利用公式(2)得到扰动梯度 g'_t 并利用 Adam 算法进行模型参数更新。

第三阶段进行模型应用,可细分为三个子步骤,该步骤结束后输出最终的分类结果及统计的分类准确率:

(1) 输入应用数据集 D_2 和隐私预算 ϵ_2 ,其中 $\epsilon_2 = \epsilon_1$;

(2) 利用公式(6)对应用数据特征进行加噪;

(3) 利用分类模型对隐私保护的应用数据进行

分类。

3.5 隐私性分析

DP-LRPCA 算法整体隐私预算 ϵ 包括 ϵ_1 和 ϵ_2 两部分,其中 ϵ_1 用于模型训练数据集的隐私保护, ϵ_2 用于应用数据集的隐私保护。隐私预算 ϵ_1 由模型在迭代过程中自适应生成,而隐私预算 ϵ_2 可以根据具体需求设置不同的值,也可以为简单起见令其取值等于 ϵ_1 。通过上述设定,使得隐私预算取值由层级相关性传播模型计算得到,且对于不同迭代步骤,隐私预算取值可能不等,以此达到自适应的目的。

第一部分在卷积神经网络模型的生成过程中实现隐私保护,利用公式(2)在梯度计算过程中添加噪声。假设 S 与 S' 是一对只有一个梯度值不同的邻近数据集,下述公式证明算法阶段 2 满足 ϵ_1 一差分隐私。

$$\begin{aligned} \frac{\Pr(\bar{F}_S)}{\Pr(\bar{F}_{S'})} &= \frac{\prod_{t=1}^{\text{iter_num}} \exp(\frac{\epsilon_t}{\Delta_F} \| \bar{g}_t \cdot |S| - \sum_{x_m \in S} g_t(x_m) \|_1)}{\prod_{t=1}^{\text{iter_num}} \exp(\frac{\epsilon_t}{\Delta_F} \| \bar{g}_t \cdot |S'| - \sum_{x'_m \in S'} g_t(x'_m) \|_1)} \\ &\leq \prod_{t=1}^{\text{iter_num}} \exp(\frac{\epsilon_t}{\Delta_F} \| \sum_{x'_m \in S'} g_t(x'_m) - \sum_{x_m \in S} g_t(x_m) \|_1) \\ &\leq \prod_{t=1}^{\text{iter_num}} \exp(\frac{\epsilon_t}{\Delta_F} \cdot \max(g_t(x_m))) \\ &\leq \prod_{t=1}^{\text{iter_num}} \exp(\epsilon_t) \\ &\leq \exp(\epsilon_1) \end{aligned} \quad (9)$$

第二部分在模型应用过程中添加隐私保护。假设应用过程中输入数据集为 L ,并且 L 与 L' 是一对邻近数据集,模型通过对图像样本的输入特征 $N_i(x_m)$ 加噪实现隐私保护,用 $\bar{h}_L$ 和 $\bar{h}_{L'}$ 分别表示加噪之后的批量数据,取 $\bar{\varphi}_i = [\sum_{x_m \in L} N_i(x_m) + \text{Lap}(\Delta_h/\epsilon_i)]$ 表示数据集 L 第 i 维扰动的输入特征之和,则下述过程证明差分隐私变换模块,即对应用数据加噪满足 ϵ_2 一差分隐私。

$$\begin{aligned} \frac{\Pr(\bar{h}_L)}{\Pr(\bar{h}_{L'})} &= \frac{\prod_{i=1}^n \exp(\frac{\epsilon_i}{\Delta_h} \| \sum_{x_m \in L} N_i(x_m) - \bar{\varphi}_i \|_1)}{\prod_{i=1}^n \exp(\frac{\epsilon_i}{\Delta_h} \| \sum_{x'_m \in L'} N_i(x'_m) - \bar{\varphi}_i \|_1)} \\ &\leq \prod_{i=1}^n \exp(\frac{\epsilon_i}{\Delta_h} \| \sum_{x_m \in L} N_i(x_m) - \sum_{x'_m \in L'} N_i(x'_m) \|_1) \\ &\leq \prod_{i=1}^n \exp(\frac{\epsilon_i}{\Delta_h} 2 \max_{x_m \in L} \| N_i(x_m) \|_1) \end{aligned}$$

$$\leq \prod_{i=1}^n \exp\left(\frac{2\epsilon_i}{\Delta_h}\right) \\ \leq \exp(\epsilon_2) \quad (10)$$

DP-LRPCA 算法可以将阶段一和二看做其子机制,因而根据串行组合定理,算法整体满足 ϵ_1 差分隐私.

4 实验结果及分析

本文使用 MNIST 数据集进行实验验证. MNIST 数据集是一个灰度手写数字图像数据集,包含训练样本 60 000 个,测试样本 10 000 个,每个样本包含一个 28×28 的灰度图像及一个分类标签. 分类准确度定义为预测正确的样本数比总的样本数. 实验环境设置为 Intel(R) Xeon(R) CPU E5-2650 v4 @ 2.20GHz, 内存 32GB, Windows Server 2008R2 Enterprise 64 位操作系统.

在上述实验环境下,我们围绕 DP-LRPCA 算法进行了 2 个实验. 其中,实验 1 是针对模型生成过程隐私保护效果的分析实验,实验 2 是在不同隐私预算下的应用数据分类性能对比实验.

实验 1 通过对比模型生成过程添加隐私保护前后的分类准确率来度量算法的可用性. 本实验分为两组,第一组以常规方式训练生成卷积神经网络,第二组结合层级相关性传播模型,并对卷积神经网络的参数更新过程进行改进,为梯度值添加自适应噪声之后用于模型参数更新,其中固定 ϵ_1 为 1,对自适应隐私预算进行累加,若超过阈值,则不对剩余迭代步骤添加噪声. 实验结果如图 2 所示.

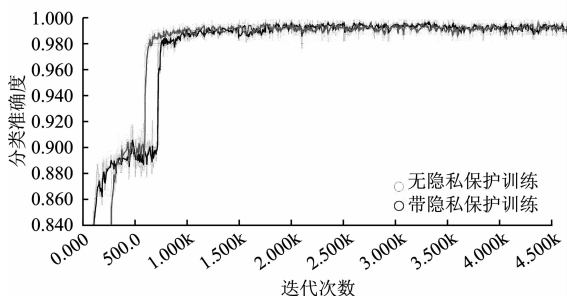


图 2 有无隐私保护的模型效果对比

由图 2 结果可知,利用 LRP 为其添加自适应噪声的隐私保护方式不会明显降低模型分类准确率,因此利用 LRP 对卷积神经网络进行自适应隐私保护具有较好的可用性.

实验 2 设置两类条件,一是选择的不同模型,原始卷积神经网络或者隐私保护卷积神经网络;二是

不同的隐私预算,组合两类条件进行应用数据的分类性能对比. 为了更好地评估输入特征扰动对分类性能的影响,实验以模型生成、模型应用步骤交替进行的方式展开,以 500 次迭代为周期进行应用分类. 为保证对比效果,实验截图展示模型生成过程,以及隐私预算 ϵ_2 分别为 0.05, 0.1, 1 时的应用数据的分类准确率. 图 3 和图 4 分别表示原始卷积神经网络和隐私保护卷积神经网络模型生成过程的分类准确率以及应用数据在不同隐私预算下的分类准确率.

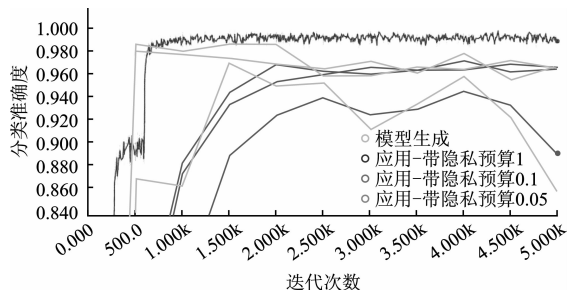


图 3 原始卷积神经网络不同隐私预算下应用数据的分类准确率

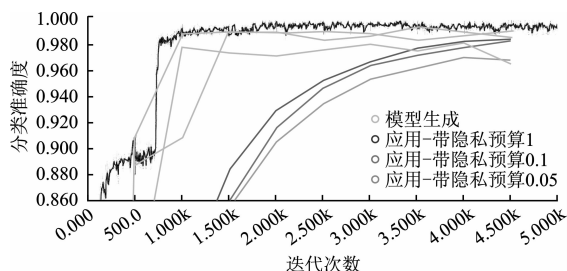


图 4 隐私保护卷积神经网络不同隐私预算下应用数据的分类准确率

图 3 在卷积神经网络未添加隐私保护的情况下进行分析,可以看到应用数据在合适的隐私预算下进行扰动,仍然能够保证 96% 以上的分类准确率,与训练数据的分类准确率平均相差 3% 左右. 而对于不同隐私预算,隐私预算越大,则隐私保护程度越小,分类准确率越高,过小的隐私预算导致扰动过大,从而导致数据可用性大大下降,符合实验设定. 图 4 在隐私保护卷积神经网络分类模型下进行应用数据分类对比,可以看到应用数据在扰动之后的分类准确率保持在 96% 以上,并且分类准确率与隐私预算的大小成正比,即隐私预算越大,分类准确率越高.

对比两组实验,通过 LRP 给图像输入特征添加自适应噪声的方式能够在隐私保护的前提下,保证

数据的分类准确率,但从时间上分析可知,扰动的应用数据在未添加隐私保护的卷积神经网络中迭代3 000次时分类准确率基本趋于稳定,而在隐私保护模型中则需要4 000次。实验结果表明,根据LRP为输入特征添加噪声的隐私保护方式具有较高的可用性,但同时为保证其分类性能,需要增加模型生成阶段的迭代次数。

5 结束语

算法通过层级相关性的相关性,并根据量化值决定对梯度值添加自适应噪声的大小和对应用数据输入特征分配的隐私预算的大小,利用自适应分配的隐私预算构建差分隐私模块实现对应用数据的扰动,实现隐私保护数据分类。实验结果表明算法在隐私保护的前提下能够保证较高的分类准确率。今后的研究内容将重点关注算法执行时间优化和解决训练过程的超隐私预算问题。

参考文献:

- [1] DWORK C. Differential privacy[C]//Proceedings of the 33rd International Colloquium on Automata, Languages, and Programming. Venice, Italy: Springer, 2006. DOI: 10.1007/11787006_1.
- [2] HAEBERLEN A, PIERCE B C, NARAYAN A. Differential privacy under fire[C]//Proceedings of the 20th USENIX Conference on Security. San Francisco: USENIX Association, 2011: 33. DOI: 10.5555/2028067.2028100.
- [3] MCSHERRY F D. Privacy integrated queries: an extensible platform for privacy-preserving data analysis[C]//Proceedings of 2009 ACM SIGMOD International Conference on Management of Data. Providence: ACM, 2009: 19-30. DOI: 10.1145/1559845.1559850.
- [4] BACH S, BINDER A, MONTAVON G, et al. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation[J]. PLoS One, 2015, 10(7): e0130140. DOI: 10.1371/journal.pone.0130140.
- [5] ZHAO J W, CHEN Y F, ZHANG W. Differential pri-

vacuity preservation in deep learning: challenges, opportunities and solutions[J]. IEEE Access, 2019(7): 48901-48911. DOI: 10.1109/ACCESS.2019.2909559.

- [6] FREDRIKSON M, LANTZ E, JHA S, et al. Privacy in pharmacogenetics: an end-to-end case study of personalized warfarin dosing[C]//Proceedings of the 23rd USENIX Conference on Security Symposium. San Diego: USENIX Association, 2014. DOI: 10.5555/2671225.2671227.
- [7] ABADI M, CHU A, GOODFELLOW I, et al. Deep learning with differential privacy[C]//Proceedings of 2016 ACM SIGSAC Conference on Computer and Communications Security. Vienna: ACM, 2016: 308-318. DOI: 10.1145/2976749.2978318.
- [8] PHAN N H, WANG Y, WU X T, et al. Differential privacy preservation for deep auto-encoders: an application of human behavior prediction[C]//Proceedings of the 30th AAAI Conference on Artificial Intelligence. Phoenix: AAAI, 2016: 1309-1316.
- [9] PHAN N H, WU X T, DOU D J. Preserving differential privacy in convolutional deep belief networks[J]. Machine Learning, 2017, 106(9-10): 1681-1704. DOI: 10.1007/s10994-017-5656-2.
- [10] XIE L Y, LIN K X, WANG S, et al. Differentially private generative adversarial network[J]. arXiv: 1802.06739, 2018.
- [11] PHAN N H, WU X T, HU H, et al. Adaptive Laplace mechanism: differential privacy preservation in deep learning[C]//Proceedings of 2017 IEEE International Conference on Data Mining. New Orleans: IEEE, 2017: 385-394. DOI: 10.1109/ICDM.2017.48.

作者简介:

侯小军 男,(1976-),硕士,工程师.研究方向为深度学习研究。

李泽华(通讯作者) 男,(1997-),硕士研究生.研究方向为深度学习、隐私保护研究。E-mail:120583174@qq.com.

李泽堃 男,(1994-),硕士研究生.研究方向为深度学习、隐私保护研究。