ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 34,No. 3,May 2019,pp. 422－431
DOI:10. 16337/j. 1004-9037. 2019. 03. 006

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

基于迁移学习的敏感数据隐私保护方法

付玉香[1]　秦永彬[1,2]　申国伟[1,2]

(1. 贵州大学计算机科学与技术学院,贵阳,550025;2. 贵州大学贵州省公共大数据重点实验室,贵阳,550025)

摘　要: 机器学习涉及一些隐含的敏感数据,当受到模型查询或模型检验等模型攻击时,可能会泄露用户隐私信息。针对上述问题,本文提出一种敏感数据隐私保护“师徒”模型PATE-T,为机器学习模型的训练数据提供强健的隐私保证。该方法以“黑盒”方式组合了由不相交敏感数据集训练得到的多个“师父”模型,这些模型直接依赖于敏感训练数据。“徒弟”由“师父”集合迁移学习得到,不能直接访问“师父”或基础参数,“徒弟”所在数据域与敏感训练数据域不同但相关。在差分隐私方面,攻击者可以查询“徒弟”,也可以检查其内部工作,但无法获取训练数据的隐私信息。实验表明,在数据集MNIST和SVHN上,本文提出的隐私保护模型达到了隐私/实用准确性的权衡,性能优越。

关键词: 差分隐私;迁移学习;模型攻击;敏感数据;隐私保护

中图分类号: TP309.2　　**文献标志码:** A

Sensitive Data Privacy Protection Method Based on Transfer Learning

Fu Yuxiang[1], Qin Yongbin[1,2], Shen Guowei[1,2]

(1. College of Computer Science and Technology, Guizhou University, Guiyang, 550025, China;2.Guizhou Provincial Key Laboratory of Public Big Data, Guizhou University, Guiyang ,550025, China)

Abstract: Machine learning involves some implicit sensitive data that may reveal user's privacy information when attacked by model attacks such as model queries or model tests. In view of the above problems, this paper proposes a sensitivity data privacy protection Mentoring model PATE-T, which provides a strong privacy guarantee for the training data for machine learning. The method combines multiple Master models trained by disjoint sensitive data sets in a black box manner, relying directly on sensitive training data. Disciple is transfer learning by Master's collection and cannot directly access Master or basic parameters. Disciple's data field is different but related to the sensitive training data field. In terms of differential privacy, an attacker can query the Disciple and check its internal work, but it cannot obtain the private information of the training data. Experiments show that the privacy protection model proposed in this paper has reached the balance of privacy/practical accuracy on the MNIST data set and SVHN data set, and the results are superior.

Key words: differential privacy; transfer learning; model attack; sensitive data; privacy protection

---

基金项目: 国家自然科学基金重大研究计划(91746116)资助项目;贵州省重大应用基础研究(黔科合JZ字[2014]2001)资助项目;贵州省科技重大专项计划(黔科合重大专项字[2017]3002)资助项目。
收稿日期: 2018-07-06;**修订日期:** 2018-08-22

引　言

机器学习(Machine learning, ML)正成为云计算时代的一种模型服务。对于数据持有人,希望能够对数据进行预测模型训练,提供机器学习框架和服务。理想情况下,将敏感数据(如病历,遗传序列等)输入到机器学习模型中训练时需要保护其隐私信息,但实际训练生成的机器学习模型难以保证。最近,利用某些隐含记忆攻击可以从ML模型中恢复敏感的训练数据。这种攻击可以直接地通过分析内部模型参数进行[1-2],也可以间接地通过反复查询不透明模型来收集数据分析攻击[3]。例如,Shokri等[1]利用会员推理攻击,根据模型的预测结果,反向推断训练模型的数据中是否包括了某些具体训练点。因此,隐私保证必须适用于最坏情况:任何隐私保护策略为了保护训练数据的隐私,应该严谨地假设攻击者可以不受限制地访问模型内部参数。为实现敏感隐私数据的可靠保护,数据脱敏技术是使用脱敏规则对某些敏感信息进行数据变形。差分隐私是经典的数据脱敏技术,添加随机噪声使敏感数据失真,同时能够保持一些数据或数据属性不变,并且保证处理后的数据在某些统计方面的性质不变,以便进行数据挖掘等操作。

如图1所示,本文采取差分隐私的数据脱敏技术,提出一种基于迁移学习隐私保护师徒模型(Private aggregation of teacher ensembles and transfer learning,PATE-T)。通过将“徒弟”的训练数据限制在“师父”投票中,并通过仔细添加随机噪声后选取最高投票。利用迁移学习把敏感的“师父”集合知识迁移到另一个非敏感数据域,进一步加强隐私保护。

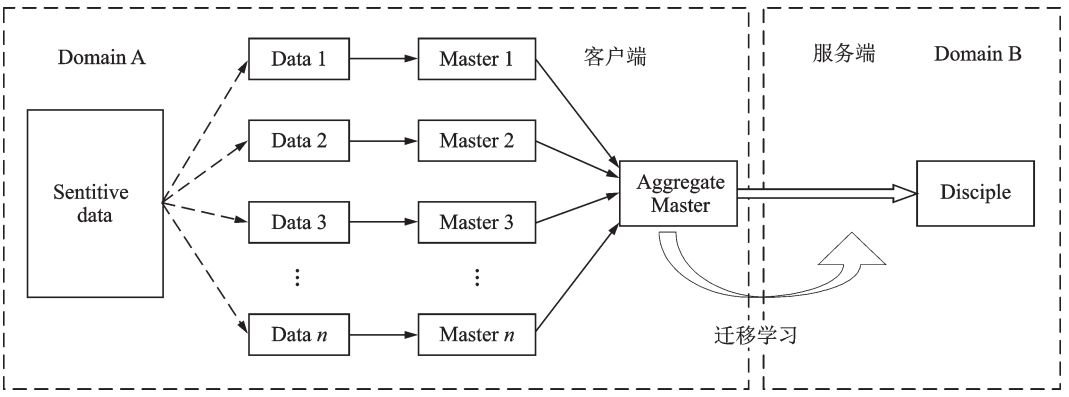

图1　PATE-T模型架构

Fig.1　PATE-T model architecture

为了确保数据的有效性,该模型包括由不相交数据子集训练生成的“师父”模型,和模仿“师父”集合的“徒弟”模型。由于所有“师父”都是在数据集的不相交子集上训练得到,当“师父”数量达到法定数量时,相应的预测源于泛化,而不是过度拟合到特定的训练点。“徒弟”在“师父”集合的总体输出上训练,确保“徒弟”不依赖任何一个敏感训练数据点。“师父”集合采用差分隐私的数据脱敏技术,保留数据在统计方面的性质,模型不会因为隐私保护而牺牲数据的有效性。本文的差分隐私学习策略仔细地添加噪声,分析和限制每个数据项的隐私影响。采用Moments accountant技术[4]动态分析“师父”嘈杂选票的敏感性,当最高投票的法定数量较大时,收紧隐私约束。

隐私保护的关键在于限制“徒弟”对“师父”的访问次数,以便“师父”能被“徒弟”进行有意义的知识表达。传统的机器学习要求领域间概率分布相同,通过传统机器学习得到的“徒弟”模型暴露于敏感数据集。因此,倘若“徒弟”模型采用传统机器学习训练,必须严格约束“徒弟”对“师父”标签查询的次数,这会在很大程度上损失“徒弟”对“师父”有效地量化,降低“徒弟”的准确性。

为了解决这个问题,本文的解决方法是利用迁移学习将源域的敏感信息知识迁移到不同但相关的

非敏感数据域中。这样设置的好处是不再强制性约束“徒弟”对“师父”的访问程度;并且能够在敏感训练数据标签很少的情况下,得到精确度很高的“徒弟”模型。在差分隐私方面,即使“徒弟”的体系结构和参数由对手进行公开或反向设计,仍保留了原始数据集上的隐私信息。根据文献[5]提出的关联思想,本文提出“徒弟”模型的关联领域自适应(Correlative domain adaptation,CDA),遵循的范例是,为了有效地导出目标域的类标签,在统计域不变嵌入空间中加强源数据与目标数据之间的关联,同时最小化标记源域的分类错误。本文提出的PATE-T方法在标准MNIST数据集[6]和SVHN数据集[7]上分别以98.46%和90.73%的分类精度取得较好的结果。

1 国内外研究现状

1.1 差分隐私

k-anonymity[8]能够保证任意一条记录与其他的$k-1$条记录不可区分,但是缺乏随机性[9],容易受到背景知识攻击和一致性攻击。Differential privacy[10](称差分隐私)给出极为严格的攻击模型的定义,对于隐私泄露风险,差分隐私给出了极为严谨的、定量化的表示和证明。

Shokri等[11]提出了一种隐私保护的分布式随机梯度下降(Stochastic gradient descent,SGD)算法,适用于非凸模型。Abadi等[4]通过Moments accountant技术对嘈杂SGD引起的隐私损失提供严格的界限,它是一种追踪隐私损失机制,允许对复杂集成机制的隐私损失进行严密的自动分析。相比之下,本文的PATE-T模型在MNIST数据集上将分类精度从97%提高到98.46%,同时将隐私约束ε从8降低到3。

Pathak等[12]首次提出了由受信任的第3方托管的全局分类器对本地分类器进行安全多方聚合。聚合通过安全协议执行,安全协议将随机组件添加到平均分类器,使所得到的聚合分类器具有隐私性,不可以推断来自本地分类器的单个数据实例。但是,这种方法不适用于具有非数值参数(如决策树)的分类器。Hamm等[13]提出了在设备上训练的一系列本地分类器知识转移到具有隐私保证的全局分类器,隐私保护模型很灵活,隐私保护和分类精度都有待提高。

Jagannathan等[14]学习了隐私保护随机森林,修改了决策树经典模型,隐私保证不是来自随机森林中由不同决策树分析的不相交训练数据集,而是来自经修改后的架构。在此基础上,Papernot等[15]提出了PATE-G方法,在敏感数据的不相关子集上训练了一组教师模型[16]。学生模型利用对抗生成网络(Generative adversarial net,GAN)[17]进行半监督学习[18]生成。本文基于这种思想对隐私保护模型进行改进,摒弃了要求训练数据和测试数据学习任务分布相同的半监督学习方法,采用迁移学习更好地保护训练数据的隐私。

1.2 迁移学习

迁移学习强调在相似但不同的领域、任务和分布间进行知识迁移,迁移学习的目标是通过一定的技术手段将源领域的知识迁移到新领域中,进而解决目标领域数据标签很少甚至没有标签的问题,不要求源领域和目标领域服从独立同分布。

迁移学习的发展从基于实例迁移,基于模型迁移,到偏重数学变换的基于特征迁移,再到深度迁移,对抗迁移。近年来,Pan等[19]提出迁移成分分析(Transfer component analysis,TCA),使用最大均值差异(Maximum mean discrepancy,MMD)[20]学习再生核希尔伯特空间(Reproducing kernel Hilbert space, RKHS)中跨领域的迁移成分。Long等[21]提出的联合分布适配(Joint distribution adaptation, JDA)使数据在降维过程中同时调整条件分布和边缘分布,并构建新的特征表示。深度适配网络(Deep adaptation network,DAN)[22]架构设计多核MMD和多层适配,将卷积神经网络推广到领域自适应场景。深度联合适配网络(Joint adaptation networks,JAN)[23]使用联合适配网络进行深度迁移学习,根据联合最

大均值差异(Joint maximum mean discrepancy,JMMD)对齐跨领域的多个特定领域层,通过联合适配网络来迁移学习。然而,用于以上迁移学习的最大均值差异MMD及其变形存在一个缺陷:需要选择适当的内核超参数,比如高斯内核的标准偏差。

Ganin等[24]提出向深度网络中加入对抗的思想。Tan等[25]提出远域迁移学习(Distant domain transfer learning,DDTL),通过选择性学习算法(Selective learning algorithm,SLA)解决目标领域数据分布与源领域数据分布完全不同的问题。Zhu等[26]使用循环一致对抗网络将一类图片转换成另一类图片。与这种方法类似,Heausser等[5]提出关联域自适应,在嵌入空间中加强源领域与目标领域之间的关联。本文的迁移学习算法CDA基于这两种方法进行改进,使目标域的分类精度更高。

2 PATE-T模型框架

如图1所示。首先"师父"集合是在互斥的敏感数据子集上训练得到;然后"徒弟"模型通过迁移学习从"师父"集合中学习得到。本节将描述数据如何分割以训练一组"师父",如何组合"师父"的预测,以及如何使用迁移学习得到"徒弟"模型。

2.1 训练"师父"队伍

数据分割:将敏感数据分为n组不相交数据集(X_n, Y_n),并分别对每组数据进行训练,摒弃了在整个数据集(X, Y)上训练单个模型的常规方法,其中,X表示输入集合,Y表示标签集合。假设n对于数据集大小和任务复杂度不算太大,得到称为"师父"的n个分类器f_i。

聚合:"师父"集合的隐私保证源于其聚合。设m为任务中的类数量。给定类$j\in 1,\cdots,m$和输入$\boldsymbol{x}$,标签计数即为分配给类j的"师父"数量:$n_j(\boldsymbol{x})=\left|\{i:i\in 1,\cdots,n, f_i(x)=j\}\right|$。如果简单地使用多数投票的方式,集合的输出可能取决于单个"师父"的投票。事实上,当两个标签的投票数最多不超过一个时,就有一个关系:如果一位"师父"作出不同的预测,总体输出就会发生变化。为了解决这个问题,本文通过添加随机噪声到投票计数n_j来引入歧义

$$f(x)=\arg\max_j\left\{n_j(\boldsymbol{x})+\mathrm{Lap}\left(\frac{1}{\varepsilon}\right)\right\} \tag{1}$$

式中:ε是一个隐私参数,$\mathrm{Lap}(b)$是以location为0和scale为b的拉普拉斯算子,参数ε影响隐私保护程度。注意,实际引入拉普拉斯机制的隐私参数ε与本文给定的ε值成反比,小的ε导致很强的隐私保证,但会降低标签的准确性。ε的选取将在本文4.2节进一步讨论。

虽然可以使用上述f进行预测,但是随着模型进行更多的预测,添加的噪声就会增加,模型将在有限数量的查询之后无效。此外,当攻击者可以访问模型内部参数时,隐私保证不成立。实际上,由于"师父"都没有考虑隐私,它们有足够的能力来保留训练数据的细节。为了解决这个问题,本文使用"师父"集合预测的标签来训练"徒弟"模型,并且将这些敏感的信息迁移学习到不同但相关的非敏感数据域,以保护敏感数据。

2.2 从"师父"集合到"徒弟"的迁移学习

使用聚合机制的输出来训练"徒弟"模型,这个"徒弟"模型的部署用来回答用户查询。隐私损失是在"徒弟"训练期间对"师父"集合标签查询相关的函数,不会随着"徒弟"模型用户查询次数而增加。使用迁移学习的方法将敏感的"师父"集合知识迁移到非敏感的"徒弟"模型,"徒弟"模型是在非敏感的公共数据集上训练得到。

训练"徒弟"模型:"徒弟"模型使用CDA方法,CDA方法是一种新的端到端的神经网络域自适应技术,基于有标记源域的统计特性来推断未标记目标域的类别标签。

在文献[27]中理论上研究域适应问题,将源域和目标域差异与各个域的统计相似度量相关联。结果表明,一个好的域适应方法应该尽可能使源和目标域相似(同化),同时尽可能地减少源域中的预测误差(歧视)。这些效应是相互对立的,源域和目标域从不同的分布中抽取。可以表述为一个成本函数

$$L = L_{\text{classification}} + L_{\text{sim}} \tag{2}$$

本文使用关联损失L_{CDA}替代相似度量(L_{sim}),最小化源域D_s上的分类错误,同时强制要求D_t与D_s具有类似统计,这可以通过加强D_t特征表示与同一类中D_s特征表示间的关联来实现[28]。

3 关联领域自适应算法

假设有两个域的数据$x_i^s \in D_s, x_i^t \in D_t$,和一个$L$层神经网络嵌入映射$\oslash: \boldsymbol{R}^{N_0} \to \boldsymbol{R}^{N_{L-1}}$,用$A_i = \oslash(x_i^s), B_j = \oslash(x_j^t)$表示源域和目标域的嵌入,对来自不同域的两个样本的相似度可以用A_i和B_j的内积$M_{ij} = \langle A_i, B_j \rangle$来表示,从嵌入$A_i$到嵌入$B_j$的转换概率形式化表示为

$$P_{ij}^{ab} = \boldsymbol{P}(B_j|A_i) = \frac{\exp(M_{ij})}{\sum_{j'}\exp(M_{ij'})} \tag{3}$$

从有标记的源域嵌入A_i开始通过无标记的目标域嵌入B,返回到另一个源域嵌入A_j的虚拟随机游走的两步往返概率表示为

$$P_{ij}^{aba} = (P^{ab}P^{ba})_{ij} \tag{4}$$

Haeusser等[28]认为,高阶往返不会提高性能。两步概率强制性要求类标签上的近似均匀分布,这可以通过称为Walker loss的交叉熵损失实现,即有

$$L_{\text{walker}} = H(T, P^{aba}) \tag{5}$$

式中:$T_{ij} = \begin{cases} 1/|A_i| & \text{class}(A_i) = \text{class}(A_j) \\ 0 & \text{其他} \end{cases}$。

这意味着同一类中的所有关联循环被迫具有相等的概率。Walker loss本身可以通过只访问容易关联的目标样本,跳过比较复杂的目标样本来最小化损失,这会导致对目标域的泛化不佳。通过调整L_{visit}可以实现以相同的概率访问每个目标样本。Visit loss由目标样本的均匀分布与任何源样本点到目标样本点的访问概率之间的交叉熵定义为

$$L_{\text{visit}} = H(V, P^{\text{visit}}) \tag{6}$$

式中:$P_j^{\text{visit}} = \sum_{x_i \in D_s} P_{ij}^{ab}$; $V_j = \frac{1}{|B|}$。

在返回映射中,进一步加强关联,增加覆盖率,Cover loss由任何目标样本开始到源样本的访问概率与源样本的均匀分布之间的交叉熵定义为

$$L_{\text{cover}} = H(P^{\text{cover}}, V) \tag{7}$$

式中:$P_i^{\text{cover}} = \sum_{x_j \in D_t} P_{ji}^{ba}$; $V_i = \frac{1}{|A|}$。

算法1 CDA算法

输入:源数据x_i^s和目标数据x_i^t,L层神经网络嵌入映射$\oslash$,权重因子$\beta_1, \beta_2, \beta_3$

输出:总体神经网络损失L

开始:

由式(4)得到从源嵌入A到目标嵌入B再返回嵌入A的两步往返概率;

利用式(5)计算随机游走损失 Walker loss;

利用式(6)计算访问损失 Visit loss;

利用式(7)计算反向随机游走损失 Cover loss;

根据式(5—7)计算源域与目标域相似嵌入的关联损失 L_{CDA},即为式(8);

计算源数据 x_i^s 的分类预测误差 $L_{\mathrm{classification}}$;

根据式(9)得到总体神经网络损失 L。

两个领域的相似嵌入之间的关联损失为

$$L_{\mathrm{CDA}}=\beta_1 L_{\mathrm{walker}}+\beta_2 L_{\mathrm{visit}}+\beta_3 L_{\mathrm{cover}} \tag{8}$$

式中 β_i 是权重因子。式(8)假定源和目标域的类分布相同,如果情况并非如此,对 L_{visit},L_{cover} 使用低权重可能会产生更好的结果。同时对网络进行训练,通过 Softmax 交叉熵损失项将源数据域的分类预测误差最小化,记为 $L_{\mathrm{classification}}$。

CDA 的总体神经网络损失为

$$L=L_{\mathrm{classification}}+L_{\mathrm{CDA}} \tag{9}$$

CDA 的关联损失强化源和目标样本的相似嵌入同化,分类损失将源数据域的预测误差最小化(歧视)。没有 L_{CDA},神经网络只能在源数据域上被传统地训练[29]。在训练期间 L_{CDA} 的加入允许合并来自不同领域的未标记数据,从而提高了分类嵌入的有效性。添加 L_{CDA} 可以使任意的神经网络进行域适配训练,这样的神经网络学习算法能够模拟源和目标域之间的分布偏移。如果 L_{CDA} 被最小化,来自源和目标域的关联嵌入在其点积上变得更相似。

4 实验分析

对 PATE-T 的评估中,集合预测标签的准确性和隐私保护性之间的权衡很大程度上取决于集合中“师父”的数量。训练一大批“师父”是必不可少的,以支持注入噪声,产生强有力的隐私保证,因此,本实验目标是在加强隐私强度的同时提高“学生”的分类精度,关键是找到合适的“师父”数量。

4.1 实验环境及数据集分析

4.1.1 “师父”模型

“师父”模型使用的卷积神经网络架构为两个带有池化、归一化的卷积层以及两个带有 ReLU 激活函数的最大全连接层。“师父”模型采用的数据集是标准 MNIST 数据集和 SVHN 数据集。本文的非隐私模式采用最先进测试结果[15]:MNIST 模型采用两个带有最大池的卷积层和一个具有 ReLUs 的完全连接层,非私人模式的测试精度为 99.18%。在此基础上,SVHN 模型另外增加两个隐藏层,非私人模式的测试精度为 92.8%。

4.1.2 “徒弟”模型

“徒弟”模型采用通用的卷积神经网络架构:$C(32,3)\rightarrow C(32,3)\rightarrow P(2)\rightarrow C(64,3)\rightarrow C(64,3)\rightarrow P(2)\rightarrow C(128,3)\rightarrow C(128,3)\rightarrow P(2)\rightarrow FC(128)$。这里 $C(n,k)$ 表示一个卷积层,n 个核的大小为 $k\times k$、步长为 1;$P(k)$ 表示一个窗口大小为 $k\times k$,步长为 1 的池化层;$FC(n)$ 表示具有 n 个输出单元的全连接层,嵌入的大小为 128。迁移学习采用的数据集是 MNIST→MNIST-M 数据集和 SVHN→MNIST 数据集。MNIST→MNIST-M:使用 MNIST 数据集作为标记源,并生成无标记的 MNIST-M 目标[30]。从彩色照片 BSDS500 数据集[31]中随机抽取背景补丁,获取每个颜色通道与 MNIST 图像差异的绝对值。与 MNIST 相比,由于两个额外颜色通道和更多细微噪音,机器识别更困难。因此,MNIST 图像的单个通道被复制 3 次以匹配 MNIST-M 图像(RGB)通道,图像大小为 28 像素×28 像素。

SVHN→MNIST:MNIST图像调整为32像素×32像素,并扩展到3个通道以匹配SVHN的形状。

4.2 聚合机制参数调整实验

式(1)呈现了拉普拉斯噪声对“师父”绩效产生的影响。聚集由不相交数据集训练的“师父”模型,向集合中注入大量随机噪声以确保隐私。此时,“师父”集合的预测依然是准确的,当$n=100$时,聚合机制的输出对于MNIST数据集的精确度是93.56%,对于SVHN数据集的精确度为87.48%,每个查询都有较低的隐私预算$\varepsilon=0.1$。

(1) 预测准确性。当其他情况相同,“师父”的数量n受限于分类任务的复杂性与可用数据之间的权衡,用分割数据集的方式训练n组“师父”,较大n导致较大的绝对差距,潜在地允许更大的噪声水平和更强的隐私保证。然而,随着n的增大,每位“师父”的训练数据随之减少,就可能降低“师父”的准确性。实验证明当$n=100$时,MNIST的个体“师父”的平均测试精度为90.45 %,SVHN的个体“师父”的平均测试精度为83.87%。

(2) 噪声聚合。对于MNIST和SVHN,考虑了具有不同数量的5组“师父”集合$n\in\{50,100,150,200,250\}$。对于每一组数据,引入不同的拉普拉斯噪声来扰乱投票数,使ε在0.01与1之间。图2显示了噪声聚合机制ε值对测试集标签准确性的影响。当逐渐降低ε值,意味着引入更多的随机噪声,隐私保护得以加强,“师父”集合的准确性很快下降。在x轴左侧较小值对应较大的噪声振幅,x轴右侧较大值对应较小的噪声振幅。

4.3 “徒弟”模型参数调整实验

通过PATE-T训练的“徒弟”模型,在SVHN→MNIST数据集上,“徒弟”分别获得5 000或10 000或15 000个训练样本,样本标签通过嘈杂的聚合机制进行标记。对于其余的21 032或16 032或11 032个样本进行评估。“徒弟”对“师父”的标签查询记为share。如图3,较多的标签查询有利于提高“徒弟”模型的准确性,share越大,“徒弟”精确度越高。在MNIST→MNIST-M数据集上,“徒弟”可以访问3 000或5 000个训练样本,训练样本通过“师父”集合进行标记,剩余的7 000或5 000个样本对性能进行评估。拉普拉斯尺度为5来保证查询隐私约束$\varepsilon=0.2$,参数的选择由2.1节所驱动。当对“师父”集合的标签查询share=5 000,“徒弟”模型精确度为98.46%,在相同的隐私保护框架下,PATE-G模型[15]中“徒弟”模型精确度为94.66%,即使是采用作者的测试结果,本实验的预测结果依然最优越。

如图4所示,n越小“徒弟”精确度越高。因为当“师父”数量较少时,个体“师父”的训练数据较多,每

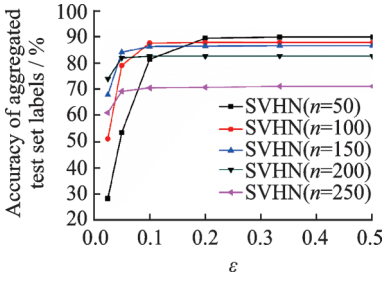

图2 噪声对聚合机制的影响

Fig.2 Effect of noise on aggregation mechanism

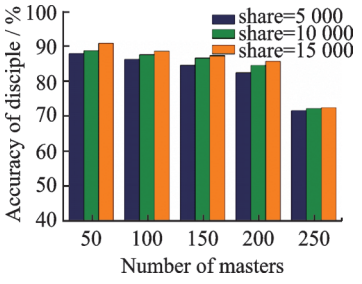

图3 不同数量的标签查询与其精确度之间的关系($\varepsilon=0.33$)

Fig.3 Relationship between different numbers of tag queries and precision ($\varepsilon=0.33$)

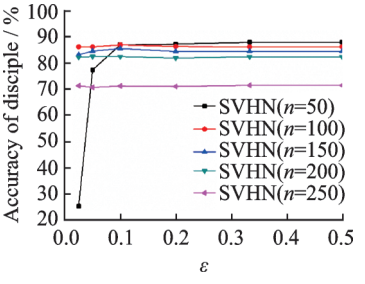

图4 噪声对“徒弟”精确度的影响

Fig.4 Effect of noise on accuracy of disciple

个“师父”都能够作出较准确的预测。此时,如果“师父”集合的噪声扰动较小,则聚合机制输出的准确性较高,“徒弟”预测精度随之较高。当$n=100$或者更大时,针对不同的噪声干扰“徒弟”的准确性趋于稳定。这种现象体现了迁移学习CDA方法的优越性,CDA方法不仅依赖于源数据域的准确性,也依赖于源数据域与目标数据域之间的相似性。在“徒弟”的训练期间,不仅有源数据域的标记数据,而且允许合并来自目标域的未标记数据,这会削弱源数据域中的噪声对“徒弟”准确性的影响。

4.4 小规模对比实验

敏感数据隐私保护方面比较前沿的技术是PATE-G模型[15],本文提出的PATE-T模型是在此基础上进行改进得到,与PATE-G模型进行对比实验。如图5,6所示,分别在MNIST和SVHN数据集上,固定“师父”的数量和“徒弟”的训练数据量,对于不同噪声干扰,PATE-T“徒弟”模型的准确率明显比PATE-G高。当噪声干扰较大时,PATE-G的准确率急剧下降,而PATE-T的性能则相对稳定,因此,当$n=50$时,PATE-T的抗噪声干扰能力优于PATE-G,且准确率高于PATE-G方法。

4.5 大规模对比实验

将结论推广到一般性,在SVHN数据集上分别进行“师父”数量为$n\in\{50,100,150,200,250\}$的5组对比实验。在此基础上,针对每组对比实验,控制用于训练“徒弟”模型的标签查询数量,将“徒弟”模型对“师父”集合的标签查询数量分别置为share$\in\{5\,000,10\,000,15\,000\}$。进一步地,在每组标签查询中,控制拉普拉斯机制的尺度参数$\varepsilon\in\{40,20,10,5,3\}$,分别对应于实验中聚合机制噪声干扰程度$\varepsilon\in\{0.025,0.05,0.1,0.2,0.33\}$。分别测试在不同“师父”数量、不同训练数据以及不同隐私保护程度下PATE-T模型与PATE-G模型的“徒弟”预测准确率,如图7所示。同理,在MNIST数据集上设置相同的实验,唯一改变的是标签查询数量share$\in\{3\,000,5\,000\}$。本文在SVHN数据集上共进行150组对比实验,在MNIST数据集上进行100组对比实验。实验证明,PATE-T模型的分类精度高于PATE-G模型的分类精度。

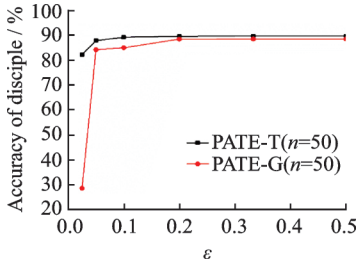

图5 不同噪声干扰ε对PATE-T,PATE-G的影响(SVHN数据集,share=10 000)

Fig.5 Effect of different noise interference ε on PATE-T, PATE-G on SVHN dataset when share=10 000

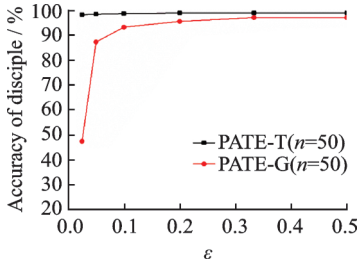

图6 不同噪声干扰ε对PATE-T,PATE-G的影响(MNIST数据集,share=5 000)

Fig.6 Effect of different noise interference ε on PATE-T, PATE-G on MNIST dataset when share=5 000

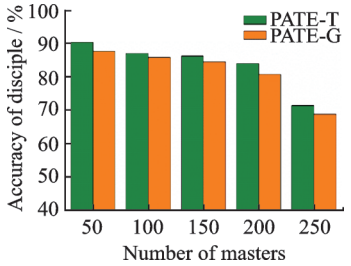

图7 PATE-T与PATE-G对比(SVHN数据集)

Fig.7 Comparison between PATE-T and PATE-G on SVHN dataset

4.6 隐私模型与非隐私模型对比

表1列出了所提供的ε值与对应“徒弟”的精确度,“徒弟”对“师父”集合标签查询数量,以及最好的非隐私模型的精确度,相应PATE-G模型精确度。噪声干扰$\varepsilon=0.2$,针对MNIST数据集,“徒弟”可以得到98.46%分类精度,与4.1节介绍的非隐私模型精确度相比只有0.72%的差距,相同条件下PATE-

表1 PATE-T模型隐私与实用

Tab. 1 Utility and privacy of the PATE-T model

Dataset	ε	Share	Non-Private baseline/%	PATE-T/%	PATE-G/%
MNIST	0.33	5 000	99.18	98.46	96.38
MNIST	0.20	5 000	99.18	98.46	94.66
SVHN	0.33	15 000	92.80	90.73	88.00
SVHN	0.20	10 000	92.80	88.54	87.16

G的分类精度只有94.66%;针对SVHN数据集,当share=15 000时,“徒弟”的准确率为90.73%,并且与4.1节介绍的非隐私模型效果相当,相应的隐私约束ε=0.33,在相同的条件下,PATE-G的测试精确度只有88.00%。

5 结束语

针对敏感训练数据的隐私保护问题,本文提出了PATE-T模型。该方法把不相交数据训练的“师父”模型进行知识聚合并迁移学习到属性可以被公开的“徒弟”模型,“徒弟”能够替代“师父”回答用户查询。由于“徒弟”与敏感的训练数据不在同一个数据域,能够强有力保证训练数据的隐私。PATE-T方法在MNIST和SVHN标准数据集精确度为98.46%和90.73%,表现显著,为用户数据提供了较好的隐私保护技术。本文提出将迁移学习运用到差分隐私保护,对训练数据进行隐私保证,对于专家和非专家人员来说都容易解释,具有较好的应用价值。目前,聚合多方隐私数据对于机器学习有很多运用价值,本文提出的PATE-T模型只是针对单源敏感训练数据进行隐私保护,后期工作将会扩展到分布式敏感训练数据隐私保护。本文的贡献是:(1)提出一种通用的机器学习策略PATE-T方法,该方法以“黑匣子”的方式为训练数据提供差分隐私,即“师父”模型和“徒弟”模型的训练方法独立于具体学习算法。(2)PATE-T保护训练数据隐私的机器模型策略,将迁移学习与隐私保护技术相结合,极大程度地提高敏感数据的隐私保护。(3)提出一种神经网络领域自适应技术CDA算法,基于有标记源域的统计特性推断无标记目标域的类别标签。

参 考 文 献:

[1] Shokri R, Stronati M, Song C, et al. Membership inference attacks against machine learning models[C]//2017 IEEE Symposium on Security and Privacy (SP). [S.l.]: IEEE, 2017: 160-176.

[2] Fredrikson M, Somesh J, Thomas R. Model inversion attacks that exploit confidence information and basic countermeasures. Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security. [S.l.]: ACM, 2015: 89-101.

[3] Zhang C, Bengio S, Hardt M, et al. Deep learning requires rethinking generalization[C]// International Conference on Learning Representation(ICLP). Toulon, France: IEEE, 2017: 262-277.

[4] 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. [S.l.]: ACM, 2016: 308-318.

[5] Haeusser P, Frerix T, Mordvintsev A, et al. Associative domain adaptation[C]//International Conference on Computer Vision (ICCV). Venice, Italy: IEEE, 2017: 2765-2773.

[6] LeCun Y, Bottou L, Bengio Y, et al. Gradient-based learning applied to document recognition[J]. Proceedings of the IEEE, 1998, 86(11): 2278-2324.

[7] Netzer Y, Wang T, Coates A, et al. Reading digits in natural images with unsupervised feature learning[C]//NIPS Workshop on Deep Learning and Unsupervised Feature Learning. Whistler, B C, Canada:NIPS, 2011: 5.

[8] Sweeney L.K-anonymity: A model for protecting privacy[J].International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 2002, 10(05): 557-570.

[9] Dwork C, Roth A. The algorithmic foundations of differential privacy[J]. Foundations and Trends in Theoretical Computer Science, 2014, 9(3/4): 211-407.

[10] Dwork C, McSherry F, Nissim K, et al. Calibrating noise to sensitivity in private data analysis[C]//Theory of Cryptography Conference. Berlin, Heidelberg: Springer, 2006: 265-284.
[11] Shokri R, Shmatikov V. Privacy-preserving deep learning[C]//Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security. [S.l.]: ACM, 2015: 1310-1321.
[12] Pathak M, Rane S, Raj B. Multiparty differential privacy via aggregation of locally trained classifiers[C]//Advances in Neural Information Processing Systems. Vancouver, British Columbia, Canada: NIPS, 2010: 1876-1884.
[13] Hamm J, Cao Y, Belkin M. Learning privately from multiparty data[C]//International Conference on Machine Learning. New York, USA:[s.n.], 2016: 555-563.
[14] Jagannathan G, Monteleoni C, Pillaipakkamnatt K. A semi-supervised learning approach to differential privacy[C]//IEEE 13th International Conference on Data Mining Workshops (ICDMW). [S.l.]: IEEE, 2013: 841-848.
[15] Papernot N, Abadi M, Erlingsson U, et al. Semi-supervised knowledge transfer for deep learning from private training data [C]// International Conference on Learning Representations. San Juan, Puerto Rico:[s.n.], 2016: 202-218.
[16] Dietterich T G. Ensemble methods in machine learning[C]//International Workshop on Multiple Classifier Systems. Berlin, Heidelberg: Springer, 2000: 1-15.
[17] Goodfellow I, Pouget-Abadie J, Mirza M, et al. Generative adversarial nets[C]//Advances in Neural Information Processing Systems. Montreal, Quebec, Canada: NIPS, 2014: 2672-2680.
[18] Chapelle O, Scholkopf B, Zien A. Semi-supervised learning[J]. IEEE Transactions on Neural Networks, 2009, 20(3): 542-542.
[19] Pan S J, Tsang I W, Kwok J T, et al. Domain adaptation via transfer component analysis[J]. IEEE Transactions on Neural Networks, 2011, 22(2): 199-210.
[20] Gretton A, Borgwardt K M, Rasch M J, et al. A kernel two-sample test[J]. Journal of Machine Learning Research, 2012, 13: 723-773.
[21] Long M, Wang J, Ding G, et al. Transfer feature learning with joint distribution adaptation[C]//Computer Vision (ICCV), 2013 IEEE International Conference on. [S.l.]: IEEE, 2013: 2200-2207.
[22] Long M, Cao Y, Wang J, et al. Learning transferable features with deep adaptation networks[C]//International Conference on Machine Learning (ICML). Lille, France:[s.n.], 2015: 97-105.
[23] Long M, Zhu H, Wang J, et al. Deep transfer learning with joint adaptation networks[C]//International Conference on Machine Learning. Sydney, NSW, Australia:[s.n.], 2017: 2208-2217.
[24] Ganin Y, Ustinova E, Ajakan H, et al. Domain-adversarial training of neural networks[J]. The Journal of Machine Learning Research, 2016, 17(1): 2096-2030.
[25] Tan B, Zhang Y, Pan S J, et al. Distant domain transfer learning[C]//Thirty-First AAAI Conference on Artificial Intelligence. San Francisco, California, USA: AAAI, 2017: 2604-2610.
[26] Zhu J Y, Park T, Isola P, et al. Unpaired image-to-image translation using cycle-consistent adversarial networks[C]//Proceedings of the IEEE International Conference on Computer Vision. Venice. Italy: IEEE, 2017: 2223-2232.
[27] Ben-David S, Blitzer J, Crammer K, et al. A theory of learning from different domains[J]. Machine Learning, 2010, 79(1/2): 151-175.
[28] Haeusser P, Mordvintsev A, Cremers D. Learning by association-a versatile semi-supervised training method for neural networks[C]//Proc IEEE Conf on Computer Vision and Pattern Recognition (CVPR). Honolulu, HI, USA:IEEE, 2017: 89-98.
[29] Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks[C]//Advances in Neural Information Processing Systems. Lake Tahoe, Nevada, United States:NIPS, 2012: 1097-1105.
[30] Ganin Y, Ustinova E, Ajakan H, et al. Domain-adversarial training of neural networks[J]. The Journal of Machine Learning Research, 2016, 17(1): 2096-2030.
[31] Arbelaez P, Maire M, Fowlkes C, et al. Contour detection and hierarchical image segmentation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2011, 33(5): 898-916.

作者简介:

付玉香(1992-),女,硕士研究生,研究方向:信息安全、隐私保护、神经网络等,E-mail: fyxfuyuxiang@163.com。

秦永彬(1980-),男,教授,博士生导师,研究方向:智能计算,机器学习和算法设计等。

申国伟(1986-),男,副教授,硕士生导师,研究方向:大数据和网络安全等。

(编辑:陈珺)