计算机科学
COMPUTER SCIENCE

自适应隐私预算分配的幸福感预测方法

罗妍婕, 李琳, 吴小华, 刘佳

引用本文

罗妍婕, 李琳, 吴小华, 刘佳. 自适应隐私预算分配的幸福感预测方法[J]. 计算机科学, 2025, 52(7): 372-378.

LUO Yanjie, LI Lin, WU Xiaohua, LIU Jia. Happiness Prediction Approach via Adaptive Privacy Budget Allocation [J]. Computer Science, 2025, 52(7): 372-378.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.240700128

自适应隐私预算分配的幸福感预测方法

罗妍婕[1]　李　琳[1]　吴小华[1]　刘　佳[2,3]
1 武汉理工大学计算机与人工智能学院　武汉 430070
2 中国科学院武汉文献情报中心　武汉 430071
3 科技大数据湖北省重点实验室　武汉 430071
(luoyanjie@whut.edu.cn)

摘　要　幸福感预测旨在通过分析个体行为、情感和社会环境等数据,预测个体生活满意度和幸福感指数。幸福感预测在线平台具有大量用户数据且存在泄露用户隐私的风险。差分隐私机器学习作为缓解该风险的有效手段,需要进一步考虑用户对不同属性的隐私需求,且现有平均分配隐私预算的差分隐私方法向模型注入了噪声,导致模型性能降低。针对上述问题,提出了一种自适应隐私预算分配的幸福感预测方法(APBA-DP)。首先根据用户的隐私偏好对属性分级,利用信息熵为属性分配个性化隐私预算;然后为幸福感预测模型建立属性映射层,基于个性化隐私预算进行差分隐私保护。在居民幸福感ESS和CGSS数据集上的实验结果表明,APBA-DP算法在一定隐私保护强度下,相比于传统差分隐私算法,准确率提升了2.3%～4.4%;同时,对其进行成员推理攻击的成功率相较于未进行差分隐私保护的模型平均降低了14.7%和12.5%。

关键词:幸福感预测;差分隐私;隐私预算
中图分类号　TP391

Happiness Prediction Approach via Adaptive Privacy Budget Allocation

LUO Yanjie[1],LI Lin[1],WU Xiaohua[1] and LIU Jia[2,3]
1 School of Computer Science and Artificial Intelligence,Wuhan University of Technology,Wuhan 430070,China
2 Wuhan Library of Chinese Academy of Science,Wuhan 430071,China
3 Hubei Key Laboratory of Big Data in Science and Technology,Wuhan 430071,China

Abstract　Happiness prediction aims to forecast individuals' life satisfaction and happiness indices by analyzing data. Online platforms for happiness prediction possess a vast amount of data,which also carries the risk of privacy breaches. Existing differential privacy machine learning methods overlook the privacy needs of different attributes. Moreover,privacy budget averaging approach injects excessive noise into the model,leading to performance degradation. To address these issues,this paper proposes a method called Adaptive Privacy Budget Allocation for Happiness Prediction(APBA-DP). Initially,attributes are graded based on users' privacy preferences,and privacy budgets are allocated using information entropy. Subsequently,happiness prediction model establishes an attribute mapping layer to ensure personalised privacy protection. Experimental results on ESS and CGSS datasets show that the accuracy of APBA-DP algorithm is improved by 2.3%～4.4% compared with the traditional differential privacy algorithms under certain privacy protection intensity. At the same time,the success rate of member inference attacks is reduced by 14.7% and 12.5% on average compared with the model without differential privacy protection.

Keywords　Happiness prediction,Differential privacy,Privacy budget

1　引言

幸福感指个体对生活满意度和主观幸福程度的感知[1]。当前,随着大数据技术在心理学的应用与发展,清华H+Lab、壹点灵等幸福感计算平台逐渐涌现。这些平台利用心理学理论和大数据分析,通过个体的基本属性、社交活动等数据,预测个体的幸福指数。这种基于线上平台的计算方法为心理学提供了一个新的研究途径。其中生态瞬时评估(Ecological Momentary Assessment,EMA)是通过智能终端进行实时数据收集的方法[2],目前,幸福感预测平台利用EMA收集的大量数据来提供在线服务。恶意攻击者调用平台的公用API,通过成员推理、模型反转攻击等隐私推断攻击

到稿日期:2024-07-22　返修日期:2024-09-18
基金项目:国家自然科学基金(62276196);科技大数据湖北省重点实验开放课题(E3KF291001)
This work was supported by the National Natural Science Foundation of China(62276196) and Hubei Key Laboratory of Big Data in Science and Technology Open Topic(E3KF291001).
通信作者:李琳(cathylilin@whut.edu.cn)

获得模型参数并模拟训练过程，进而侵犯模型训练数据的隐私[3]。因此，如何在防止敏感信息泄露的情况下预测幸福感成为亟需解决的问题。

差分隐私[4](Differential Privacy，DP)是机器学习中的一种隐私保护方法，通过在数据收集或模型训练期间添加受控噪声来保护训练数据的隐私。然而，当前的差分隐私方法往往采用统一处理敏感属性的方式均分隐私预算，这种做法不仅会影响模型的预测准确性，而且无法实现敏感属性的个性化保护。幸福感预测涉及收集和处理个人数据，如收入、健康状况、社交关系等，而这些数据的敏感程度因人而异。因此，本文提出的自适应差分隐私算法根据被调查者的隐私偏好为不同属性分配个性化的隐私预算。例如，对于特别重视收入隐私的人，算法会为收入数据分配更多的隐私预算；而对于那些更关心健康隐私的人，健康数据将获得更多的隐私保护。基于上述问题，本文提出了一种自适应差分隐私的幸福感预测方法，根据被调查者的隐私偏好为不同属性分配个性化的隐私预算，从而有效减少隐私预算的浪费。此方法旨在满足数据提供者的个性化隐私需求，同时，减小噪声对模型准确性产生的负面影响。本文的主要贡献为：

1)对幸福感数据集的属性进行敏感和非敏感两大主要级别的划分，以满足调查对象对个人数据属性隐私的差异化需求；

2)利用敏感属性的信息熵设计了一种隐私预算分配策略，该方法有效缓解了平均分配隐私预算导致的模型准确度显著下降和敏感属性保护不足的问题；

3)理论证明所提算法满足差分隐私，并在幸福感数据集上进行实验评估，结果表明该算法能够兼顾模型的可用性与安全性，满足个性化需求。

本文旨在研究如何通过自适应隐私预算分配来预测个体的幸福感。第1章介绍了隐私保护在幸福感预测中的重要性及研究动机；第2章回顾了当前机器学习的幸福感计算方法和差分隐私模型训练相关算法；第3章介绍了自适应隐私预算分配的幸福感预测方法的具体流程，包括具体的隐私预算分配策略和差分隐私随机梯度下降的应用；第4章分析了模型预测准确率及隐私保护效果；最后总结全文并展望未来。

2 相关工作

2.1 幸福感计算方法

现有的幸福感研究常利用机器学习方法对个体幸福指数进行预测，并挖掘出影响幸福指数的因素。表1对部分利用机器学习进行幸福感计算的研究进行总结。

一些学者关注于传统机器学习方法，例如 Garaigordobil 等[5]从健康和社交能力方面，利用多元回归分析(Multiple Regression Analysis，MRA)探究影响幸福感等级的重要变量；Saputri 等[6]利用信息增益技术和支持向量机(Support Vector Machine，SVM)以及信息增益等降维技术预测国民幸福感；You 等[7]利用中国群众的幸福指数及相关调查数据集，采用线性回归、决策树(Decision TREE，DT)等方法分析与幸福指数相关的特征，并预测幸福感。Fan 等[8]利用模糊特征生成方法提出了 FF-SVM，FF-CatBoost 等算法，提高幸福感预测的准确性；Zhang 等[9]通过多元回归和贪婪算法设计情绪表征检测(Emotion Representatives Detection，ERD)方法进行社交网络的幸福感分析。上述传统机器学习方法通常只能捕捉线性或简单的非线性关系，无法很好地衡量多个因素之间复杂的相互作用。

随着深度学习的发展，一些学者应用深度学习进行复杂数据的幸福感分析。Chaipornkaew 等[10]在多层感知机(Multi-Layer Perceptron，MLP)的基础上采用过采样和欠采样技术建立幸福感预测模型来研究员工的满意度和快乐指数。文献[11]设计了一种抽样方法，使用 Shapley 值和深度神经网络来阐明个体因素对幸福预测的贡献。Li 等[12]利用卷积神经网络(Convolutional Neural Network，CNN)提取面部特征，结合长短期记忆(Long Short Term Memory，LSTM)网络构建基于多元回归序列输入的幸福水平预测(Sequential Inputs via Multiple Regressions，SIVML)。Cerekovic 等[13]提出了 Face 模型，依靠 LSTM 来编码群体结构，利用从人脸中获得的信息，从快乐程度和图像空间分布预测幸福感。Ding 等[14]提出了一种基于图卷积网络(Graph Convolution Network，GCN)的大学教师幸福预测方法，通过综合分析学术创新、工作满意度等因素来预测教师的幸福趋势。Li 等[15]基于深度神经网络(DNN)和 Shapley 值，提出了一种在幸福感计算中利用联盟博弈论计算因子间相互作用的解决方案。Kumar 等[16]提出了一种基于 Transformer 的情感检测系统，使用上下文相关特征来更好地从文本中捕捉用户的心理状态。上述研究在采用深度学习方法预测幸福感时，忽略了数据隐私泄露的风险。

表1 幸福感计算相关研究

Table 1 Related work on happiness computing

分类	方法	基础模型
传统机器学习	DR-SVM[6]，FF-SVM[8]	SVM
	LR[5]，UML[7]，ERD[9]	MAR
	UML[7]，FF-CatBoost[8]	DT
深度学习	SMOTE-MLP[10]	MLP
	SIVML[12]	CNN
	SIVML[12]，Face[13]	LSTM
	BERT-Base-FT[16]	Transformer

2.2 差分隐私模型训练

基于大数据的幸福感研究依靠大量的训练数据作为模型的输入，当模型的训练结果以接口的形式呈现在网络平台上，那些涉及调查对象隐私的训练数据就很可能遭受到攻击。表2总结了模型训练中运用到的部分差分隐私保护方法。

表2 差分隐私模型训练相关研究

Table 2 Related work on differential privacy model training

分类	方法	对属性分配个性化隐私预算
梯度扰动	DP-SGD[17]	×
	DPAGD-CNN[18]	×
模型参数扰动	UDP[19]	×
目标函数扰动	AdLM[21]	√(输入数据加噪)

部分学者在机器学习模型和结果中结合差分隐私保护训练数据。例如，Ruan 等[17]优化了差分隐私随机梯度下降(Differential Privacy Stochastic Gradient Descent，DP-SGD)

算法。Huang 等[18]提出了卷积神经网络差分隐私自适应梯度下降(Differential Privacy Adaptive Gradient Descent CNN, DPAGD-CNN)算法,在每次迭代中自适应地分配不同的隐私预算。Wei 等[19]提出了一种用户级差分隐私(User-Level Differential Privacy, UDP)算法,该算法在共享模型上传到服务器之前添加噪声。PCAFMG-DPLR 算法[20]以扰动后的目标函数最小化为目标,求得最优模型的参数。此外,Phan 等[21]开发了一种自适应拉普拉斯机制(Adaptive Laplace Mechanism, AdLM),通过扰动目标函数、输入数据以及参数对模型进行隐私保护。文献[22]提出了基于功能机制的深度差分隐私自编码器(Deep Private Autoencoder, DPA),通过噪声化目标函数保护自编码器输出中的数据敏感信息。

然而,以上差分隐私与模型结合的方法大多未充分考虑数据属性敏感度的差异,而默认对不同属性的保护程度是相等的。以性别和年龄为例,用户更倾向于保护年龄信息而不是性别信息,且性别属性仅具有两个属性值,因此相较于年龄属性,携带的信息量较少。属性所含信息量不同,对攻击者推断目标对象隐私信息的贡献程度存在差异,这意味着为了降低隐私泄露风险,需要重点保护敏感属性。

3 自适应隐私预算分配的幸福感预测方法

3.1 差分隐私基本概念

差分隐私保证在数据集中对一条记录进行增删等操作,查询返回的输出无明显差异,因此攻击者不能推断出目标记录是否在发布的数据集中[4]。

定义 1(差分隐私) 对于任意两个相邻数据集 D_1 和 D_2,它们之间最多相差一条记录。$Range(M)$ 是随机算法 M 的取值范围,若算法 M 在数据集 D_1 和 D_2 上的任意输出 $S\in Range(M)$,满足:

$$P(M(D_1)\in S)\leqslant e^{\varepsilon}\times \Pr(M(D_2)\in S)+\delta \tag{1}$$

则称算法 M 满足(ε,δ)-DP。ε 为隐私预算参数,代表了差分隐私保护水平,其值越小,隐私保护级别越高;δ 指代一个常数,代表可以容忍违反严格差分隐私的概率,特别地,当 δ 为 0 时,算法满足 ε-DP。

定义 2(全局敏感度)[21] 设查询函数 $f:D\rightarrow\mathcal{R}^n$,$f$ 的全局敏感度定义如下:

$$\Delta f=\max_{D_1,D_2}\| f(D_1)-f(D_2)\| \tag{2}$$

其中,D_1 和 D_2 表示任意两个相邻数据集,全局敏感度 Δf 由查询函数 f 决定。

定义 3(Laplace 机制)[21] 对任意数据集和查询函数 $f:D\rightarrow\mathcal{R}^n$,若算法 M 的输出满足:

$$M(D)=f(D)+Lap(\Delta f/\varepsilon) \tag{3}$$

则算法 M 符合 ε-DP。$Lap(\Delta f/\varepsilon)$ 表示添加的噪声量,噪声量与 Δf 成正比,与 ε 成反比。

定义 4(高斯机制)[17] 对任意数据集和查询函数 $f:D\rightarrow\mathcal{R}^n$,$\boldsymbol{N}\sim\mathcal{N}(0,c\Delta f/\varepsilon)$,若算法 M 的输出结果满足:

$$M(D)=f(D)+\boldsymbol{N} \tag{4}$$

则算法提供了(ε,δ)-DP,$c\geqslant\sqrt{2\ln(1.25/\delta)}$。

3.2 问题描述

本文关注利用机器学习方法研究幸福感所面临的训练数据隐私泄露问题,重点在于,针对隐私等级差异为不同属性合理地制定隐私保护方案,以平衡预测模型的准确性与隐私保护强度。

如图 1 所示,研究框架中涉及一个可信的数据收集方(例如社会调查机构或心理研究机构),该方利用深度学习技术对用户数据进行分析,并将所得模型提供给第三方服务提供商,以构建幸福感计算平台。然而,这些服务提供商并非完全可信,同时恶意攻击者有可能通过调用平台的公共 API 利用成员关系推理和模型反演等攻击手段,获取模型参数并模拟训练过程,从而导致严重的隐私泄露问题。在幸福感预测的数据中,被调查者对不同属性的隐私需求存在显著差异。例如,收入数据通常高度敏感,具有较高的隐私保护等级;年龄和性别等属性的敏感度较低,隐私保护等级也偏低。

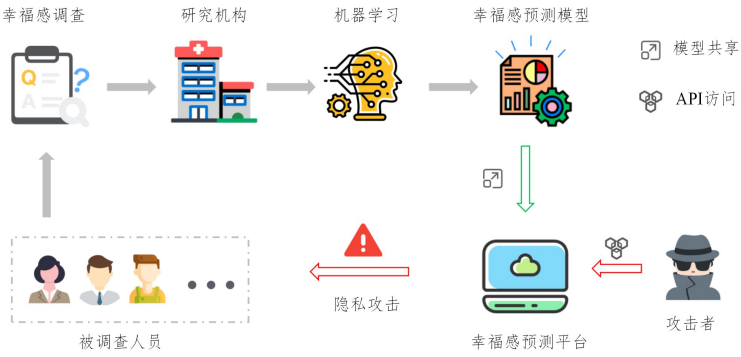


图 1 在线幸福感计算隐私泄露

Fig. 1 Privacy leakage in online happiness computing

3.3 自适应隐私预算分配的差分隐私算法

根据隐私等级,针对幸福感数据属性,进行不同的隐私预算分配。如图 2 所示,自适应差分隐私算法可分解为以下 3 个基本步骤:

1)根据用户需求,对数据集属性进行分类,将其划分为非敏感属性和敏感属性。对于敏感属性,根据其信息量的大小,分配相应的隐私预算,非敏感属性则保持一致的隐私预算。

2)引入差分隐私属性映射层。该层的神经元以全连接的结构组成,根据自适应隐私预算对仿射变换进行扰动。

3)将 MLP, CNN, LSTM 和 Transformer 作为基础模型进行幸福感预测时,采用 DP-SGD 算法进行训练,确保整个模型满足-DP。

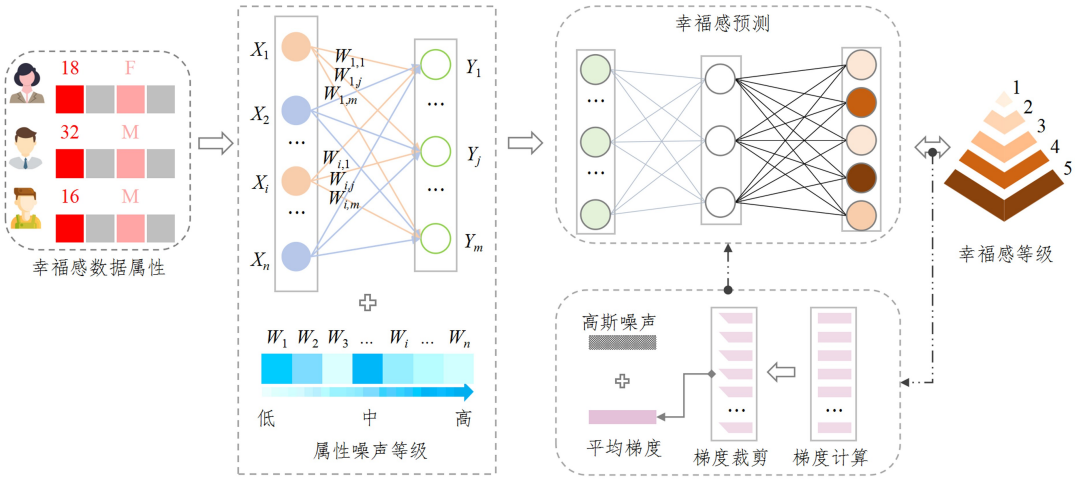


图 2 自适应隐私预算分配的幸福感预测方法框架

Fig. 2 Framework of happiness prediction with adaptive privacy budget allocation

3.3.1 自适应属性隐私预算分配策略

基于用户隐私需求，将数据集中的属性划分为敏感属性和非敏感属性。考虑到每个敏感属性所包含的信息量存在差异，对其进行隐私保护时需采取不同的保护强度。为此，引入信息熵对敏感属性进行分级处理，以确保隐私保护措施能够有效地应对不同级别的信息泄露风险。APBA-DP 算法伪代码如算法 1 所示。

算法 1 APBA-DP 算法

输入：训练样本数据 $D\in\mathcal{R}^{n\times m}$，$\{(x_i,y_i)\}$

输出：属性映射层参数 w^0，幸福感预测模型参数 w^1

1. 随机初始化参数 w^0，w^1

损失函数 $L(\theta)=\frac{1}{|D|}\sum_i L(\theta;x,y)$

属性映射层输出 $f(w^0,x)$

2. for j←1 to $|A_s|$ do //计算 A_s中属性信息熵 H

$H_i=-\sum P(D^i_{A^j_s})\log(D^i_{A^j_s})$

3. for m←1 to M do //计算属性的隐私预算 e

if m in $A_s$then

$e_m=\frac{\varepsilon_1}{2-e^{-(H_m-\min(H))}}$

else then

$e_m=\varepsilon_1$//非敏感属性统一分配

4. for t←1 to T do //差分隐私保护训练

5. $w^0=w^0/\max(1,\|w^0\|_1/C_w)$ //裁剪参数

6. for j ←1 to M do //添加噪声并更新参数

$w^0_j=w^0_j+\mathrm{Lap}(\frac{T\Delta S_D}{e_j})$

7. for i ←1 to |D|//d do

8. 随机挑选大小 d 的样本集

9. //计算梯度值

$g_t(x_i,y_i)\leftarrow\nabla\theta_t(L(w^1;f(w^0,x_i),y_i))$

10. //裁剪梯度：

$\tilde{g}_t(x_i,y_i)\leftarrow\tilde{g}_t(x_i,y_i)/\max(1,\|g_t(x_i,y_i)\|_2/C_g)$

11. //梯度聚合并添加噪声

$\tilde{g}_t(x_i,y_i)\leftarrow\tilde{g}_t(x_i,y_i)+\frac{1}{d}\mathcal{N}\left(\frac{c\Delta S_1}{\varepsilon_2}\right)$

12. $w^{1,t}\leftarrow w^{1,t-1}-\eta_t\tilde{g}_t$ //更新模型参数

在算法 1 中，为属性分配的隐私预算不再采取均分策略，而是以信息熵为权重分配隐私预算。设定数据的属性集为 U，根据用户需求划分为敏感属性集和非敏感属性集，对应的索引分别表示为 A_s和 A_n。对非敏感属性采用统一的隐私预算 ε_1。步骤 2 对于敏感属性集中的每个属性计算其信息熵：

$$H_m=-\sum p(D_{A_s^m,q})\log p(D_{A_s^m,q}) \tag{5}$$

其中，H_m 为敏感属性集 A_s中第 m 个属性的信息熵，$p(D_{A_s^m,q})$代表第 m 个敏感属性取值为 $D_{A_s^m,q}$的概率。

设定为敏感属性分配的最大隐私预算为 ε_1，步骤 3 对每个敏感属性分配的隐私预算为：

$$e_m=\frac{\varepsilon_1}{2-e^{-(H_m-\min(H))}} \tag{6}$$

3.3.2 自适应隐私预算分配策略下的属性映射

步骤 4 为 APBA-DP 算法的差分隐私保护训练。首先构建一个差分隐私属性映射层，在仿射变换中注入自适应拉普拉斯噪声来扰动每个与输入数据直接关联的参数，以保护特定属性的隐私。结合敏感属性隐私预算分配策略，对敏感级别不同的属性实现差异化隐私保护。

定义属性映射层的训练过程为：

$$S\rightarrow w_T=\frac{1}{|D|}\sum_{i=1}^{|D|}\arg\min F(w_T,D_j) \tag{7}$$

步骤 5 通过裁剪技术确保属性映射层的训练参数小于裁剪阈值 C_w。根据定义 2 计算 S_D 的敏感度：

$$\begin{aligned}\Delta S_D&=\max\|S_D-S_D'\|\\&=\max\|\frac{1}{|D|}\sum_{i=1}^{|D|}\arg\min F(w_T,D_i)-\\&\quad\frac{1}{|D|}\sum_{i=1}^{|D|}\arg\min F(w_T,D'_i)\|\\&=\frac{2C}{|D|}\end{aligned} \tag{8}$$

其中，D 和 D'是两个相邻数据集，D_j'是 D_j'中的第 j 个样本。

步骤 6 向属性映射层的第 j 个参数 w_j^0 注入拉普拉斯噪声：

$$w_j^0=w_j^0+Lap(\frac{\Delta S_D}{e_j}) \tag{9}$$

数据集 D 经过属性映射层的仿射变换 $f(\bar{W}^0,D)$可改写为：

$$f(\bar{W}^0,D)=\sum_{j=1}^{M}\left(\sum_{x_i\in D}x_{ij}\left(W_j^0+Lap\left(\frac{\Delta S_D}{e_j}\right)\right)\right)$$

$$=x\sum_{j=1}^{M}\left(W_j^0+Lap\left(\frac{\Delta S_D}{e_j}\right)\right) \tag{10}$$

$$\phi_j=W_j^0+Lap\left(\frac{\Delta S_D}{e_j}\right) \tag{11}$$

其中，ϕ_j 为扰动后的属性映射层的第 j 个参数。由此，数据集 D 经过属性映射层得到一个特定输出的概率为：

$$\Pr(f(\bar{W}^0,D))=\prod_{j=1}^{M}\exp\left(\frac{e_j\parallel x_jW_j^0-x_j\phi_j\parallel}{\Delta S_D}\right) \tag{12}$$

相邻数据集 D 和 D' 经过属性映射层的结果满足：

$$\begin{aligned}&\frac{\Pr(f(\bar{W}^0,D))}{\Pr(f(\bar{W}^0,D'))}\\&=\frac{\prod_{j=1}^{M}\exp\left(\frac{e_j\parallel x_jW_j^0-x_j\phi_j\parallel}{\Delta S_D}\right)}{\prod_{j=1}^{M}\exp\left(\frac{e_j\parallel x'_jW_j^0-x'_j\phi_j\parallel}{\Delta S_D}\right)}\\&\leqslant\prod_{j=1}^{M}\exp\left(\frac{e_j}{\Delta S_D}\parallel x_jW_j^0-x_j\phi_j-x'_jW_j^0+x'_j\phi_j\parallel\right)\\&\leqslant\prod_{j=1}^{M}\exp\left(\frac{e_j}{\Delta S_D}\parallel(x_j-x'_j)(W_j^0-\phi_j)\parallel\right)\\&\leqslant\prod_{j=1}^{M}\exp\left(\frac{e_j}{\Delta S_D}\parallel(W_j^0-\phi_j)\parallel\right)\\&\leqslant\prod_{j=1}^{M}\exp\left(\frac{e_j}{M\times\Delta S_D}\right)\\&\leqslant\exp(\varepsilon_1)\end{aligned}$$

因此，属性映射层的仿射变换满足 ε_1-DP。

3.3.3 差分隐私随机梯度下降

目前在深度学习领域，应用最广泛的差分隐私算法是差分隐私随机梯度下降[17]（DP-SGD）。该算法能够在保护数据隐私的同时，实现模型的优化。为了在基础模型训练过程中保护每个样本的隐私，算法 1 中的步骤 8－步骤 12 采用了 DP-SGD 算法来训练基础模型。每轮训练随机选择一个批量为 d 的训练样本，然后根据损失函数和训练标签计算模型的梯度，对模型梯度进行裁剪并引入高斯噪声。由差分隐私中的高斯机制可得，基础模型总体满足 ε_2-DP[17]。

定理 1（差分隐私串行组合定理[23]） 给定数据集 D 和 D 上一组随机独立算法 $M_1(D),M_2(D),\cdots,M_m(D)$，其中每个算法 $M_i(D)$ 满足 ε_i-差分隐私，则组合算法 $M=\{M_1,M_2,\cdots,M_m\}$ 满足 $\sum_{i=1}^{m}\varepsilon_i$-差分隐私。由定理 1 可得，算法 1 的整体训练过程满足 $(\varepsilon_1+\varepsilon_2,\delta)$-DP。

4 实验结果与分析

4.1 数据集和评价指标

本文选取了 2015 年中国综合社会调查（Chinese General Social Survey，CGSS[1]）数据集和 2018 年欧洲社会调查（European Social Survey，ESS[2]）数据集进行研究。CGSS 数据集包含超过 8000 个样本，每个样本含有个人信息（性别和职业）、经济情况（家庭状况和保险）等 140 个属性。ESS 是一项跨国调查，涵盖了 30 多个欧洲国家。该调查涉及多个主题，包括但不限于公众信任、政治兴趣以及幸福、健康等方面。ESS 数据集中包含近 5 万条数据记录，每条记录包含 102 个属性。在幸福感预测的多分类任务中，采用准确率（Accuracy，Acc）和加权平均 F_1 作为评价指标。在二分类的成员推理攻击任务中，采用准确率作为攻击效果的评价指标。

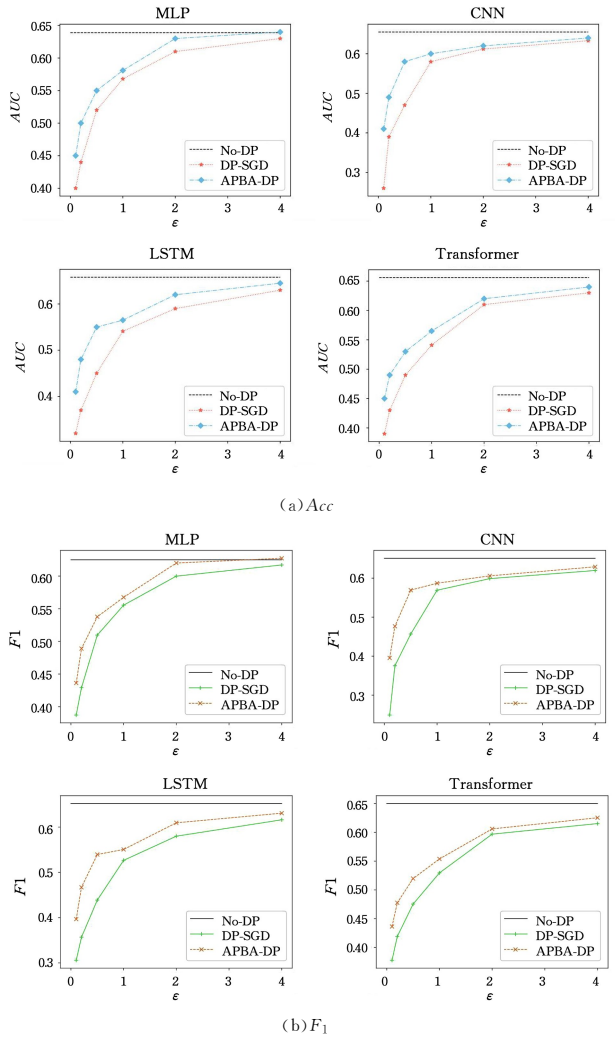
4.2 实验设置

对两个数据集统一设置训练集、测试集和验证集的比例为 7∶2∶1，并选取前 10%的属性作为敏感属性。实验选用的幸福感预测基础模型包括表 1 中的 MLP，CNN，LSTM 和 Transformer。训练批量大小为 64，训练轮次为 30。为公平比较 APBA-DP 算法与 DP-SGD 算法在基础模型上的性能和隐私保护效果，固定高斯机制的参数 $\delta=1\times10^{-4}$，$c=2$，梯度裁剪阈值 $C_g=10$。在采用 APBA-DP 算法训练模型时，属性映射层参数的裁剪阈值为 $C_w=5$，为敏感属性分配的最大隐私预算 $\varepsilon_1=0.4$。

4.3 实验结果及分析

4.3.1 模型预测准确率分析

图 3 用 Acc 和加权 F_1 两项指标衡量了 APBA-DP 和 DP-SGD 两种差分隐私算法在 ESS 数据集上训练基础模型达到的效果。



(a) Acc

(b) F_1

图 3 ESS 数据集上不同隐私预算下模型的评估效果（CGSS 效果类似）

Fig. 3 Model evaluation under different privacy budgets on ESS dataset(CGSS similar)

[1] http://www.cnsda.org/index.php?r=projects/view&id=6207244

[2] https://ess.sikt.no/en/?tab=overview

黑线表示未采用差分隐私训练模型的基线水平。相同基础模型在 CGSS 数据集上的性能与 ESS 数据集上的性能类似。实验结果表明，对于较小的 ε，相比于采用均分隐私预算的 DP-SGD 算法训练整个模型，APBA-DP 算法训练模型具有更高的性能。DP-SGD 在每次梯度更新中引入了噪声，导致在处理高维度数据或复杂任务时，模型精度下降较为明显。APBA-DP 通过动态调整隐私参数，向模型中注入较少的噪声，从而有效降低了对模型准确度的不利影响。尤其值得注意的是，在采用 CNN 作为基础模型训练 ESS 数据集时，当整体模型满足的隐私保护强度 $\varepsilon \approx 0.1$ 时，采用 DP-SGD 算法达到的测试准确率仅为 26.1%，而 APBA-DP 算法则能达到 41.2%。

4.3.2 隐私保护效果分析

本文通过成员推理攻击[3]（Membership Inference Attack，MIA）评估使用 APBA-DP 算法训练模型与使用 DP-SGD 训练整个模型的隐私保护效果。成员推理攻击是一种重要的隐私威胁，可以推断出某个特定样本是否被用于训练模型。为了进行这项评估，实验设计了一个线性攻击模型，并使用其对不同隐私保护算法进行测试。攻击模型的准确率（Acc）反映了隐私保护算法的隐私效能：高攻击成功率（即高准确率）表明隐私保护算法的保护效果较差，反之则表示其保护效果良好，也就是说，如果攻击者能够成功地推断出某个样本是否在训练集中，那么模型的隐私保护能力就被认为是较低的，反之，如果攻击者的成功率较低，则说明模型较好地保护了训练数据的隐私。实验中，总体隐私预算设定为 1，使用成员推理攻击对幸福感预测模型进行了攻击。攻击结果如表 3 所列。使用了 3 种不同的训练方法进行对比：未采用任何差分隐私算法的模型（No-DP）、采用 APBA-DP 算法训练的模型，以及采用 DP-SGD 算法训练的模型。结果表明，未采用差分隐私算法的模型（No-DP）的成员推理攻击成功率最高，与预期一致。相比之下，采用差分隐私算法训练的模型表现出更低的成员推理攻击成功率，表明差分隐私算法有效增强了模型的隐私保护能力。特别是，使用 APBA-DP 算法训练的模型成功地降低了成员推理攻击的成功率，在精确度高于 DP-SGD 的情况下，抵御成员推理攻击的效果与 DP-SGD 算法相似。这说明 APBA-DP 算法在保证较高的模型预测能力的同时，在保护训练数据隐私方面具有良好的效果。

表 3 成员推理攻击效果对比（总体隐私预算为 1）

Table 3 Comparison of membership inference attack effectiveness($\varepsilon=1$)

（%）

Fundation Model	DP Method	ESS			CGSS		
		Acc ↑	*F1* ↑	*Attack-acc* ↓	*Acc* ↑	*F1* ↑	*Attack-acc* ↓
MLP	No-DP[10]	**0.639**	**0.625**	0.661	**0.604**	**0.482**	0.614
	+DP-SGD[17]	0.568	0.544	**0.568**	0.517	0.435	**0.518**
	+APBA-DP(ours)	0.581	0.565	0.570	0.531	0.452	0.521
CNN	No-DP[12]	**0.655**	**0.651**	0.689	**0.620**	**0.535**	0.624
	+DP-SGD[17]	0.589	0.557	0.551	0.540	0.471	0.537
	+APBA-DP(ours)	0.609	0.594	**0.541**	0.552	0.485	**0.532**
LSTM	No-DP[13]	**0.658**	**0.652**	0.653	**0.625**	**0.564**	0.598
	+DP-SGD[17]	0.525	0.485	**0.565**	0.521	0.491	**0.516**
	+APBA-DP(ours)	0.542	0.521	0.570	0.533	0.527	0.521
Transformer	No-DP[16]	**0.656**	**0.650**	0.692	**0.615**	**0.556**	0.633
	+DP-SGD[17]	0.541	0.525	0.620	0.539	0.516	0.591
	+APBA-DP(ours)	0.565	0.552	**0.615**	0.544	0.526	**0.587**

结束语 本文针对幸福感预测的隐私保护问题，提出了一种自适应隐私预算分配的差分隐私算法（APBA-DP），以缓解传统差分隐私方法中忽视用户隐私需求差异和注入大量噪声导致精确度下降的问题。本文通过实验验证了 APBA-DP 算法在幸福感数据集上的可行性和有效性。实验结果表明，APBA-DP 算法成功缓解了平均分配隐私预算可能导致的模型准确度下降和敏感属性保护不足的问题，为幸福感预测提供了一种有效的隐私保护方案。未来的研究将进一步优化 APBA-DP 算法，寻找更精细的隐私预算分配策略，以减少潜在的隐私泄露风险；同时，探索将本文方法扩展到其他领域的可能。

参 考 文 献

[1] ZHANG T W. Relationship between Income Level, Income Gap and Subjective Happiness: Empirical Analysis Based on CGSS Data of Six Provinces in 2017[J]. Areal Research and Development, 2021, 40(6): 31-36.

[2] SHI Y W, CHEN T Z, DU J. The application of EMA method in mental and psychological digital medicine[J]. Chinese Journal of Nervous and Mental Diseases, 2023, 49(8): 503-508.

[3] RIGAKI M, GARCIA S. A survey of privacy attacks in machine learning[J]. ACM Computing Surveys, 2023, 56(4): 1-34.

[4] GAO M, ZUO F, WANG G. Efficient Differential Privacy Federated Learning Mechanism for Intelligent Selection of Optimal Privacy Protection Levels[C]// International Conference on Web Information Systems and Applications. Cham: Springer, 2022: 603-614.

[5] GARAIGORDOBIL M. Predictor variables of happiness and its connection with risk and protective factors for health[J]. Frontiers in Psychology, 2015, 6: 133172.

[6] SAPUTRI T R D, LEE S W. A study of cross-national differences in Happiness factors using machine learning approach[J]. International Journal of Software Engineering and Knowledge Engineering, 2015, 25(9/10): 1699-1702.

[7] YOU L. Utilizing machine learning to predict happiness index [C]// 2021 2nd International Conference on E-Commerce and Internet Technology (ECIT). IEEE, 2021: 233-238.

[8] FAN Z, GOU J, WENG S. A Novel Fuzzy Feature Generation Approach for Happiness Prediction[J]. IEEE Transactions on Emerging Topics in Computational Intelligence, 2024, 8(2): 1595-1608.

[9] ZHANG X, LI W, HUANG H, et al. Predicting happiness state based on emotion representative mining in online social networks[C]// Pacific-Asia Conference on Knowledge Discovery and Data Mining. 2017: 381-394.

[10] CHAIPORNKAEW P, PREXAWANPRASUT T. A prediction model for human happiness using machine learning techniques [C]// 2019 5th International Conference on Science in Information Technology(ICSITech). IEEE, 2019: 33-37.

[11] LI L, WU X H, KONG M, et al. Towards the Quantitative Interpretability Analysis of Citizens Happiness Prediction[C]// Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence(IJCAI). IEEE, 2022: 5094-5100.

[12] LI J, ROY S, FENG J, et al. Happiness level prediction with sequential inputs via multiple regressions[C]// ACM International Conference on MultimodalInteraction. 2016: 487-493.

[13] CEREKOVIC A. A deep look into group happiness prediction from images[C]// Proceedings of the 18th ACM International Conference on Multimodal Interaction. 2016: 437-444.

[14] DING H. Prediction analysis of college Teachers' happiness based on the graph convolutional network[J]. Mathematical Problems in Engineering, 2022, 2022: 1-9.

[15] LI L, WU X H, KONG M, et al. Quantitatively Interpreting Residents Happiness Prediction by Considering Factor-Factor Interactions[J]. IEEE Transactions on Computational Social Systems, 2024, 11(1): 1402-1414.

[16] KUMAR A, CAMBRIA E, TRUEMAN T E. Transformer-based bidirectional encoder representations for emotion detection from text[C]// 2021 IEEE Symposium Series on Compu- tational Intelligence(SSCI). IEEE, 2021: 1-6.

[17] RUAN W, XU M, FANG W, et al. Private, efficient, and accurate: Protecting models trained by multi-party learning with differential privacy[C]// 2023 IEEE Symposium on Security and Privacy. IEEE, 2023: 1926-1943.

[18] HUANG X, GUAN J, ZHANG B, et al. Differentially private convolutional neural networks with adaptive gradient descent [C]// 2019 IEEE Fourth International Conference on Data Science in Cyberspace(DSC). IEEE, 2019: 642-648.

[19] WEI K, LI J, DING M, et al. User-level privacy-preserving federated learning: Analysis and performance optimization [J]. IEEE Transactions on Mobile Computing, 2021, 21(9): 3388-3401.

[20] LI K J, HU X X, CHEN Y, et al. Differential Privacy Linear Regression Algorithm Based on Principal Component Analysis and-Functional Mechanism[J]. Computer Science, 2023, 50(8): 342-351.

[21] PHAN N H, WU X, HU H, et al. Adaptive laplace mechanism: Differential privacy preservation in deep learning[C]// 2017 IEEE International Conference on Data Mining(ICDM). IEEE, 2017: 385-394.

[22] PHAN N H, WANG Y, WU X, et al. Differential privacy preservation for deep auto-encoders: an application of human behavior prediction[C]// Proceedings of the AAAI Conference on Artificial Intelligence. AAAI, 2016: 1309-1316.

[23] ZHANG X, ZHANG X, Wang Q Y. DP-IMKP: Data Publishing Protection Method for Personalized Differential Privacy[J]. Computer Engineering and Applications, 2023, 59(10): 288-298.

LUO Yanjie, born in 2000, postgraduate. Her main research interests include data mining and machine learning.

LI Lin, born in 1977, Ph.D, professor, Ph.D supervisor, is a member of CCF (No. 34840M). Her main research interests include multi-modal machine learning, information retrieval and recommender systems.

(责任编辑:喻藜)