基于差分隐私的数据脱敏技术研究

李思慧[1]，戴明超[1]，蔡伍洲[2]
（1. 武警吉林省总队，吉林　长春　130000；2. 武警部队作战勤务局，北京　100000）

摘　要：随着人工智能和大数据技术的发展，全球数据产业规模和数据储量呈爆发式增长。在挖掘数据价值的同时，确保数据安全已成为亟需解决的关键问题。数据脱敏技术通过预先设定的规则和算法，对敏感数据进行变换，去除数据中的敏感信息，可防止敏感数据被非法访问、获取，又可以减少对整体数据集挖掘利用的影响，实现了保持数据可用性的同时，保护用户的隐私数据。针对神经网络预测模型中的数据隐私保护问题，利用差分隐私技术中的 Laplace 机制对 Adult 数据集进行脱敏，并在神经网络预测模型中进行验证，对比原始数据、差分隐私脱敏数据及其他脱敏技术数据生成模型的预测效果，结果表明，经差分隐私技术处理后的数据，既保证了数据隐私，又实现了数据的有效利用。

关键词：数据脱敏；差分隐私；Laplace 机制
中图分类号：TP309　　**文献标识码：**A　　**DOI：**10.19358/j.issn.2097-1788.2025.02.006
引用格式：李思慧，戴明超，蔡伍洲. 基于差分隐私的数据脱敏技术研究[J]. 网络安全与数据治理，2025，44(2)：39-43.

Research on data desensitization technology based on differential privacy

Li Sihui[1], Dai Mingchao[1], Cai Wuzhou[2]
(1. Jilin Armed Police Corps, Changchun 130000, China;
2. Combat Service Bureau, People's Armed Police, Beijing 100000, China)

Abstract: With the development of artificial intelligence and big data technology, the global data industry is experiencing explosive growth in scale and data reserves. Ensuring data security while mining its value has become a critical issue that urgently needs to be solved. Data desensitization technology transforms sensitive data with pre-set rules and algorithms, removing sensitive information from the data, preventing illegal access and acquisition of sensitive data, and reducing the impact on the overall data set mining and utilization, achieving privacy protection while maintaining data availability. This article addresses the issue of data privacy protection in neural network prediction models. The Laplace mechanism in differential privacy technology is used to desensitize the Adult data set and validated in the neural network prediction model. Comparing the predictive performance of the original data, differential privacy-sanitized data, and models generated from other desensitization techniques data, the results show that the data processed by differential privacy technology not only ensures data privacy but also achieves effective utilization of the data.

Key words: data desensitization; differential privacy; Laplace mechanism

0　引言

当前，人工智能、大模型、大数据技术飞速发展，数据是各项技术构建的关键基础资源，全球数据产业正在呈爆发式增长。据国际数据公司（IDC）预测，2018～2025 年，全球数据量将从 33 ZB 猛增至 175 ZB，而根据工业和信息化部相关预测，2021～2025 年，我国的大数据产业规模将从 1.3 万亿元突破至 3 万亿元，数据已然成为推动经济社会发展最重要的基础生产要素之一[1]。数据资源被充分利用的同时，数据安全问题也日益凸显，数据被非法获取事件频频发生，给企业和个人带来了巨大损失。因此，在挖掘数据价值的同时，确保数据安全，已成为亟需解决的关键问题。

传统的数据安全解决方案大多关注于数据的存储和传输，在对数据进行挖掘利用时，仍然需要具有敏感信

息的原始数据，数据非法窃取者可通过身份攻击、属性攻击、存在性攻击和概率知识攻击等，推断出个体敏感信息[2]。数据脱敏技术是通过对数据进行一定处理来保护隐私的技术，其目的是在保留输入数据的统计特征以及可用性的同时，保护数据的隐私和安全[3]。差分隐私技术是数据脱敏技术的一种，该技术提供了一种隐私保护方法，旨在向原始数据注入噪声或扰动，实现在保护个体数据隐私的同时，完成对数据的挖掘利用[4]。

差分隐私技术在国外研究较早，且技术日趋成熟。2006 年，Dwork 等人[5]首次提出了差分隐私保护方法，该方法通过向原始数据添加服从特定分布的噪声，用以保护敏感数据，解决了传统数据匿名脱敏技术无法抵抗背景知识攻击的问题。2016 年，Abadi 等人[6]提出了具有差分隐私的深度学习算法，分析了差分隐私在深度学习框架内的隐私成本，在保护数据隐私的同时，训练出有效的深度学习模型。2019 年，Holohan 等人[7]设计了 IBM 差分隐私库，用于 Python 编程语言中研究、实验和开发差分隐私应用程序。2023 年，Holohan[8]又提出了差分隐私随机数生成器和种子算法，实现了在差分算法和结果中进行测试和错误修复，为差分隐私算法选择提供了有利帮助。

近年来，国内的差分隐私技术研究也取得了丰硕成果。2009 年，袁进良[9]设计了统一的差分隐私联邦学习平台，扩展了传统的隐私预算组合定理，实现了随时间不断更新的可用预算，解决了差分隐私的强隐私和联邦系统的高吞吐难兼顾问题。2023 年，张连福[10]提出了一种基于同态加密与差分隐私的隐私保护联邦学习方案，利用多种防护措施实现了隐私防护范围覆盖联邦学习全生命周期。同年，张旭[11]提出一种兼顾安全防御和隐私保护的分布式学习系统，该系统实现隐私保护的同时，提升了训练模型的准确性。随着差分隐私技术的不断迭代发展，其在数据隐私保护领域得到越来越多的应用。

本文探讨基于差分隐私的数据脱敏方法，对数据集进行清洗整理后，利用 Laplace 机制对敏感数据进行处理，利用神经网络模型分别对未脱敏数据和脱敏后的数据进行训练和预测，对比原始数据、差分隐私脱敏数据及其他脱敏技术数据生成模型的预测效果，为神经网络预测模型的数据隐私保护问题提供解决方案。

1　数据脱敏技术

数据脱敏技术主要包括同态加密技术、数据匿名技术和差分隐私技术。3 种技术适用于不同的数据类型及应用场景。

同态加密是一种特殊的加密方法，它允许加密数据在不暴露明文的情况下进行计算，得到的结果仍然是加密的结果[12]。数据匿名技术利用数据隐蔽、泛化、混排、扰动等技术，对数据进行隐藏、模糊或替换处理，解除用户隐私属性与身份属性的关联。差分隐私技术根据特定的规则和数学方法，向原始数据集中的数据添加随机噪声，降低攻击者识别个体信息的机会，确保输出的数据信息受单条数据记录的影响始终低于某个阈值，从而使攻击者无法根据输出数据的变化判断单条记录的增删或修改，同时，添加噪声的数据对整体数据集研究分析影响较小。差分隐私的实现机制主要包括拉普拉斯机制、指数机制和随机响应机制等[13]。3 种数据脱敏技术适用场景及优缺点对比如表 1 所示。

表 1　数据脱敏技术对比

脱敏技术	优点	缺点	适用场景
同态加密	支持复杂计算	计算密集	云计算
	提高安全性	通信成本高	医疗保健
	保护隐私	效率低	金融机构
		技术复杂	区块链
数据匿名	简单易行	数据失真	数据共享
	保护隐私	评估困难	市场调研
			数据分析
差分隐私	严格定义隐私保护	模型准确性降低	政府数据分析
	易于集成	参数选择困难	企业数据分析
	保护单个样本隐私	累积噪声	人工智能领域

本文研究数据脱敏技术在机器学习算法中的应用，考虑到机器学习算法参数多，算法复杂，训练数据集数据量大，且同态加密技术不能进行非线性运算，因其无法实现大量数据的密文快速处理[14]，数据匿名技术会导致数据失真，因此，选用差分隐私算法对敏感数据进行脱敏。

1.1　差分隐私定义

1.1.1　邻近数据集

设数据集 D、D'具有相同的属性结构，两者的差记作 $D\Delta D'$，$|D\Delta D'|$表示对称差的数量，若 $|D\Delta D'|=1$，则称 D、D'为邻近数据集。

1.1.2　全局敏感度

对一个查询函数 f：$D\rightarrow \boldsymbol{R}^n$，其中 D 为一个数据集，$\boldsymbol{R}^n$为查询返回的 n 维实数结果向量，在任意一对邻近数据集 D、D'上的全局敏感度定义为：

$$GS_{f(D)}=\max\|f(D)-f(D')\| \tag{1}$$

式中，$\|f(D)-f(D')\|$表示 $f(D)$ 与 $f(D')$ 间的曼哈

顿距离。

1.1.3 差分隐私

设有隐私算法 M，Range（M）为隐私算法 M 输出结果的取值范围，P 为 M 所有可能输出集合构成的概率，若算法 M 在邻近数据集 D、D' 上输出结果 S_M（$S_M \in$ Range（M））满足式（2），则称算法 M 为数据集提供了 ε-差分隐私保护。

$$P[M(D) \in S_M] \leqslant e^{\varepsilon} \times P[M(D') \in S_M] \quad (2)$$

式中，ε 越小隐私保护度越高，数据可用性越低；ε 越大隐私保护度越低，数据可用性越高。$\varepsilon = 0$ 时，M 针对 D、D'的输出概率完全相同。

1.2 Laplace 机制

Laplace 机制是将服从 Laplace 分布的随机噪声添加到原始数据中来实现差分隐私保护。给定数据集 D，设有函数 f：$D \rightarrow \boldsymbol{R}^n$，其敏感度为 Δf，那么随机算法 M（D）为：

$$M(D) = f(D) + \mathrm{Lap}\left(\frac{\Delta f}{\varepsilon}\right) \quad (3)$$

M（D）满足 ε-差分隐私保护，则该方法服从 Laplace 分布，分布参数为$\frac{\Delta f}{\varepsilon}$。

为了实现 ε-差分隐私保护，Laplace 机制会对原始数据添加服从 Laplace 分布的随机噪声，该随机噪声的概率密度函数表示为：

$$f(x \mid u, b) = \frac{1}{2b} e^{-\frac{|x-u|}{b}} \quad (4)$$

其中，u 为位置参数，决定了噪声的中心位置；b 为尺度参数，决定了噪声的强度，b 越大，加入的随机扰动就越大。

Laplace 机制可以根据需要调整其尺度参数 b 从而控制噪声的大小，具有灵活性。通过加入服从 Laplace 随机扰动的噪声，原始数据就会被扰动，从而增加了数据的隐私性。Laplace 机制在差分隐私的框架下，能够为数据提供一定级别的隐私保护，并降低随机噪声对模型预测准确性的影响。

2 差分隐私技术在神经网络模型中的应用

为了验证差分隐私算法的有效性，本文建立了差分隐私算法验证模型，主要流程如图 1 所示。

首先利用 Laplace 噪声对 Adult[15] 数据集进行脱敏处理，将原始数据和脱敏数据分别建立训练集、验证集、测试集，利用神经网络预测模型对数据集进行训练、验证和测试，对比分析原始数据、拉普拉斯机制脱敏数据和其他机制脱敏数据在神经网络预测模型中的预测效果。考虑到经同态加密机制处理的脱敏数据无法实现神经网络模型激活函数 sigmoid 函数的计算，本文仅对比基于

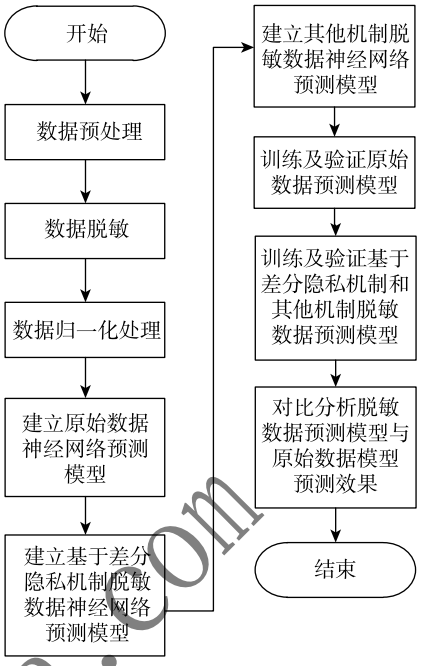

图 1 差分隐私数据脱敏技术验证流程图

Laplace 机制和基于数据匿名机制的脱敏数据在神经网络预测模型中的预测效果。

2.1 数据预处理

本文采用 Adult 数据集进行差分隐私数据脱敏技术验证。Adult 数据集是一个广泛使用的机器学习数据集，用于预测个体的年收入是否超过 50k 美元。该数据集来自 1994 年美国的人口普查数据，经过处理适用于机器学习分类任务。数据集包含多个特征，这些特征描述了每个个体年龄、性别、职业等不同方面信息，数据集的目标变量是一个二元标签，指示个体的年收入是否超过 50k 美元。

在利用数据集前，首先对数据进行预处理，删除有缺失的数据项，去除重复数据，统一数据格式，删除空白和特殊字符，采用独热编码技术将离散型文本数据转换为数值型数据。

2.2 数据脱敏

引入 Laplace 算子，设置 $u = 0$，定义差分隐私参数 ε，其取值范围为［0，1］，根据实际应用的需求和隐私风险，加入调整参数 λ，形成新的隐私参数为 $\lambda \times \varepsilon$，为每个数据集添加基于 Laplace 分布的噪声，检测添加噪声后的数据合理性，对超出边界合理值的数据设置为边界值，整合脱敏后的数据集。

2.3 模型预测

对原始数据和脱敏后数据进行归一化处理。对数据集按 6 ∶2 ∶2 划分建立训练集、验证集、测试集。因为

此数据集预测结果为二分类，本文创建了两层感知机网络，并对结果进行 softmax 归一化处理。训练及验证每经过一个 epoch 就进行一次损失比较，当损失值较上次迭代更小时，保存模型，直至迭代结束。下载在训练过程保存的最好模型进行预测并保存结果。

2.4 对比脱敏效果

对原始数据、拉普拉斯机制脱敏数据和其他机制脱敏数据在神经网络模型上完成训练和测试后，对结果进行对比，评估脱敏后的数据对神经网络预测模型的影响。设实际收入超过 50k 且预测收入超过 50k 为 TP，实际收入超过 50k 且预测收入低于 50k 为 FN，实际收入低于 50k 且预测收入超过 50k 为 FP，实际收入低于 50k 且预测收入低于 50k 为 TN。本文的评估度量将采用准确率（Precision，P）、召回率（Recall，R）和查全率（准确率和召回率的调和平均值）$F1$ 值作为评估指标。公式如下：

$$P = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}} \tag{5}$$

$$R = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{6}$$

$$F1 = \frac{P \times R}{P + R} \times 2 \tag{7}$$

3 实验分析

3.1 软硬件环境

实验软硬件环境配置如下：

（1）CPU：AMD Ryzen7 5800H 3.20 GHz；

（2）内存：16.0 GB；

（3）显卡：RTX 3060 6 GB；

（4）硬盘：500 GB；

（5）操作系统：Windows10 64 位；

（6）应用软件：Python 3.8、PyCharm、Anaconda3。

3.2 实验运行结果及结果分析

实验运行结果如表 2 所示。

表 2 脱敏数据与原始数据运行结果比较（%）

数据类型	原始数据	差分隐私脱敏数据	匿名脱敏数据
Precision	83.9	82.5	79.1
Recall	75.8	73.3	71.3
$F1$	80.7	77.6	75.2

由表 2 可见，基于差分隐私机制的脱敏数据预测模型的准确率相比原始数据预测模型略有下降，从 83.9% 降至 82.5%，召回率从 75.8% 降至 73.3%，说明经过脱敏处理后，预测模型在预测为正例的个体样本中呈现更多的假阳性。查全率从 80.7% 降至 77.6%，这是由于数据经过脱敏处理后，预测模型的准确率和召回率均有所下降，导致查全率有所下降。从表 2 还可以看出，基于差分隐私机制脱敏数据训练的神经网络模型预测准确率和召回率均高于基于匿名机制脱敏数据训练的神经网络模型。又由于基于同态加密的脱敏数据无法进行激活函数 sigmoid 函数的运算，因此，在对神经网络预测模型的数据进行脱敏处理时，选择基于差分隐私机制的脱敏技术更佳。

4 结论

本文针对神经网络预测模型中的数据隐私保护问题，利用基于 Laplace 机制的差分隐私脱敏算法对 Adult 数据集进行脱敏处理，并对比原始数据、差分隐私脱敏数据及其他脱敏技术数据在神经网络预测模型中的预测效果。实验结果表明，基于差分隐私机制脱敏数据训练的神经网络模型综合指标与原数据训练的神经网络模型相比有所下降，但高于基于匿名机制脱敏数据训练的模型。经差分隐私技术处理后的数据，既保证了数据隐私，又实现了数据的有效利用，本文研究为基于差分隐私机制的数据脱敏在机器学习中的应用提供了一定参考。未来工作中，将研究如何设置自适应参数，以提高使用脱敏数据预测模型的准确率；研究如何实现同态加密、数据匿名和差分隐私技术在机器学习中的混合运用。

参考文献

[1] 沈传年，徐彦婷. 数据脱敏技术研究及展望 [J]. 信息安全与通信保密，2023（2）：105－116.

[2] 臧帅，朱友文. 匿名数据集隐私保护效果度量机制 [J]. 网络空间安全科学学报，2024（3）：67－76.

[3] 刘建智. 机器学习中基于同态加密的隐私保护技术研究 [D]. 成都：电子科技大学，2024.

[4] 蔡彬彬. 基于差分隐私的多维数据脱敏方法研究 [J]. 信息记录材料，2024，25（5）：160－162.

[5] DWORK C，MCSHERRY F，NISSIM K，et al. Calibrating noise to sensitivity in private data analysis [J]. Lecture Notes in Computer Science，2006（3876）：265－284.

[6] 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.

[7] HOLOHAN N，BRAGHIN S，AONGHUSA P M，et al. Diffprivlib：the IBM differential privacy library [J]. ArXiv：1907.02444，2019.

[8] HOLOHAN N. Random number generators and seeding for differential privacy [J]. ArXiv：2307.03543，2023.

[9] 袁进良. 面向移动终端的资源高效联邦学习方法研究 [D]. 北京：北京邮电大学，2019.
[10] 张连福. 联邦学习隐私保护关键问题研究 [D]. 南昌：江西财经大学，2023.
[11] 张旭. 基于差分隐私机制的数据隐私保护的研究 [D]. 南京：南京邮电大学，2023.
[12] 王伟艳. 基于同态加密的安全多方计算协议的优化与应用 [D]. 太原：中北大学，2023.
[13] 李建华，银鹰，李思源，等. 大数据安全与隐私计算技术综述 [J/OL]. 网络空间安全科学学报，1－15 [2024－12－18]. https://doi.org/10.20172/j.issn.2097－3136.240601.
[14] 戴怡然，张江，向斌武，等. 全同态加密技术的研究现状及发展路线综述 [J]. 电子与信息学报，2024，46 (5)：1774－1786.
[15] Adult：UCI machine learning repository [DB/OL]. [2024－10－18]. https://doi.org/10.24432/C5XW20.

（收稿日期：2024－12－18）

作者简介：

李思慧（1985－），女，硕士，工程师，主要研究方向：大数据、人工智能。

戴明超（1989－），男，硕士，助理工程师，主要研究方向：大数据、人工智能。

蔡伍洲（1989－），男，本科，工程师，主要研究方向：大数据、人工智能。

（上接第 38 页）

[19] GOSWAMI P, MADAN S. Privacy preserving data publishing and data anonymization approaches: a review [C]//2017 International Conference on Computing, Communication and Automation (ICCCA), 2017: 139－142.
[20] MARQUES J F, BERNARDINO J. Analysis of data anonymization techniques [C]//KEOD, 2020: 235－241.
[21] BERTINO E, OOI B C, YANG Y J, et al. Privacy and ownership preserving of outsourced medical data [C]//21st International Conference on Data Engineering (ICDE' 05), 2005: 521－532.
[22] NAKAMURA Y, NAKATSUKA Y, NISHI H. Novel method to watermark anonymized data for data publishing [J]. IEICE Transactions on Information and Systems, 2017, 100 (8): 1671－1679.
[23] NAKAMURA Y, NISHI H. Digital watermarking for anonymized data with low information loss [J]. IEEE Access, 2021, 9: 130570－130585.
[24] BORDEL B, ALCARRIA R, ROBLES T, et al. Data authentication and anonymization in IoT scenarios and future 5G networks using chaotic digital watermarking [J]. IEEE Access, 2021, 9: 22378－22398.
[25] YU J, YUAN S G, YUAN Y L, et al. A k-anonymity-based robust watermarking scheme for relational database [C]//Science of Cyber Security: 4th International Conference, SciSec 2022, 2022: 557－573.

（收稿日期：2024－12－01）

作者简介：

李程（1994－），通信作者，男，硕士，助理工程师，主要研究方向：数据安全、隐私保护。E-mail: licheng2021@alumni.pku.edu.cn。

张国栋（1995－），男，本科，助理工程师，主要研究方向：信息安全。

版权声明

凡《网络安全与数据治理》录用的文章，如作者没有关于汇编权、翻译权、印刷权及电子版的复制权、信息网络传播权与发行权等版权的特殊声明，即视作该文章署名作者同意将该文章的汇编权、翻译权、印刷权及电子版的复制权、信息网络传播权与发行权授予本刊，本刊有权授权本刊合作数据库、合作媒体等合作伙伴使用。同时，本刊支付的稿酬已包含上述使用的费用，特此声明。

《网络安全与数据治理》编辑部