基于数据特征的成员推理增强攻击方法

牛俊[1]，沈括[2]，王月琮[3]，韩雪雪[3]，张嘉伟[1]，李兴华[1]，张玉清[1,2,4]
（1.西安电子科技大学网络与信息安全学院，陕西 西安 710126；2.西安电子科技大学杭州研究院，浙江 杭州 311231；
3.西安电子科技大学通信工程学院，陕西 西安 710071；4.国家计算机网络入侵防范中心（中国科学院大学），北京 101408）

摘　要：针对现有成员推理攻击（MIA）方法只在单一实例或场景下检测攻击的有效性，导致检测结果不准确并误导攻击者判断这一问题，首先构造了多个测试场景并联合多个评估指标，设计了一种成员推理攻击有效性检测算法，发现现有攻击方法的假阳率（FPR）最高可达100%。在此基础上，根据数据特征和攻击原理，提出了一种基于数据特征的成员推理增强攻击方法，以降低攻击的假阳率并提高攻击性能。实验结果表明，所提方法在4个基准数据集上，可将15个现有成员推理攻击的假阳率从100%降低到0，将精确率从58.27%提高到100%，将成员增益从32.37%提高到100%，进一步增强了攻击的有效性。

关键词：机器学习；成员推理攻击；数据安全；数据特征
中图分类号：TP391
文献标志码：A
DOI：10.11959/j.issn.1000–436x.2025162

Enhanced membership inference attack method leveraging data characteristics

NIU Jun[1], SHEN Kuo[2], WANG Yuecong[3], HAN Xuexue[3],
ZHANG Jiawei[1], LI Xinghua[1], ZHANG Yuqing[1,2,4]
1. School of Cyber Engineering, Xidian University, Xi'an 710126, China
2. School of Hangzhou Institute of Technology, Xidian University, Hangzhou 311231, China
3. School of Telecommunications Engineering, Xidian University, Xi'an 710071, China
4. National Computer Network Intrusion Protection Center (University of Chinese Academy of Sciences), Beijing 101408, China

Abstract: Aiming at the problem that the existing membership inference attack (MIA) methods only present generic testing results and had not sufficiently evaluated the effectiveness of such attacks, which could mislead attackers, multiple test scenarios were first constructed and various evaluation metrics were combined to propose a membership inference attack validity detection algorithm. It was revealed that the false positive rate (FPR) of current attack methods could reach as high as 100%. Based on data characteristics and attack principles, an enhanced membership inference attack method was subsequently proposed to reduce the false positive rate and improve attack effectiveness. Experimental results on four benchmark datasets demonstrate that the FPR of 15 existing membership inference attacks is reduced from 100% to 0, while the attack precision is improved from 58.27% to 100%, and the membership advantage is increased from 32.37% to 100%, further enhancing the effectiveness of the attack.

Keywords: machine learning, membership inference attack, data security, data characteristics

收稿日期：2025–05–13；修回日期：2025–09–15
通信作者：张玉清，zhangyq@nipc.org.cn
基金项目：国家自然科学基金资助项目（No.U2336203, No.62125205, No.U23A20303）；国家重点研发计划基金资助项目（No.2023YFB3106400, No.2023QY1202）；北京市自然科学基金资助项目（No.L242015, No.4242031）；西电概念验证基金资助项目（No.GNYZ2024QC009）；中央高校基本科研业务费专项基金资助项目（No.ZYTS25066）；陕西省自然科学基础研究计划基金资助项目（No.2024JC-YBMS-544）
Foundation Items：The National Natural Science Foundation of China (No.U2336203, No.62125205, No.U23A20303), The National Key Research and Development Program of China (No. 2023YFB3106400, No. 2023QY1202), Beijing Natural Science Foundation (No.L242015, No.4242031), The Innovation Fund of Xidian University (No.GNYZ2024QC009), The Fundamental Research Funds for the Central Universities (No.ZYTS25066), The Natural Science Basic Research Program of Shaanxi (No.2024JC-YBMS-544)

0　引言

算力、算法和数据的不断进步，促进了机器学习（ML, machine learning）的快速发展，进而推动了人工智能（AI, artificial intelligence）的繁荣与创新。机器学习作为AI的一种实现方式，在图像识别、自然语言处理、数据挖掘[1]、欺诈识别[2]、医学分析等领域取得了巨大进展。然而，机器学习服务（MLaaS, machine learning as a service）模型的训练需要大量数据，而这些数据涉及隐私信息，如位置轨迹、医疗和金融消费等，其严重威胁着机器学习的数据安全。同时，即使机器学习模型和算法是可信的，机器学习模型的输出（如信任分数向量）也很容易包含训练数据的隐私信息，从而导致其易遭受各种隐私攻击，如模型提取攻击、模型投毒攻击、后门攻击、对抗样本攻击、成员推理攻击（MIA, membership inference attack）[3-4]以及模型版权确权等。因此，关注和研究机器学习中的数据安全具有至关重要的意义。

成员推理攻击[3,5-7]主要推测一个数据样本是否在目标模型的训练集中，极大地威胁着训练数据的隐私安全。尽管目前已有大量成员推理攻击方法[3-7]依赖目标模型的输出概率、真实标签、影子模型和数据集的差异性等信息，来捕捉成员关系并提升攻击性能。然而，现有工作只给出了一般的测试结果，并未充分检测这些攻击方法的有效性和实用性，进而导致错误的推测结果误导攻击者并造成巨大的经济损失。为解决上述问题，本文首先根据数据特征构造了多个测试场景（TS, test scenarios），提出了一种成员推理攻击有效性检测算法，其先对目标数据进行分类并构造了4种测试场景，接着计算不同测试场景下各个攻击方法的攻击效果，并与预先设置的假阳率（FPR, false positive rate）阈值和成员增益（MA, membership advantage）阈值进行比较来判断攻击的有效性。检测出现有MIA存在较高的假阳率（如100%），而且不同数据集和类别的假阳率也不同。此外，根据攻击时使用的目标数据、数据特征和攻击原理，提出了一种基于数据特征的成员推理增强攻击方法，其主要根据攻击者能容忍的最小假阳率以及数据样本间的距离，对样本进行摘选和替换。在4个基准数据集上的实验发现，利用本文方法可极大地提升15个现有先进MIA的攻击性能。例如，将假阳率降低到0，将精确率和成员增益等评估指标的数值提高到100%，有效增强了攻击性能。本文的主要贡献包括以下3个方面。

1) 提出了一种成员推理攻击有效性检测算法，发现现有方法存在较高的假阳率，最高可达100%，而且不同数据集和类别具有不同的假阳率。

2) 提出了一种基于数据特征的成员推理增强攻击方法，其根据目标数据特征（如模型输出概率或数据样本间的距离）及攻击原理，来挑选合适的目标数据以实施攻击，极大地增强了攻击效果。

3) 对比15种现有MIA表明，本文方法在4个基准数据集上可极大增强各个攻击的攻击效果。例如，将攻击假阳率降低到0，将精确率和成员增益提高到100%。

1　相关工作

1.1　成员推理攻击

目标ML模型F被描述为在某个目标训练数据集D_{train}上训练的ML模型。MIA旨在确定数据点x是否为目标训练数据集D_{train}中的成员。MIA包括三类：基于二元分类器的MIA利用影子技术训练一个二元攻击模型；基于评估机制的MIA根据预先定义的成员评估机制（如模型输出阈值）来实施攻击；基于数据集差异的MIA不需要训练影子模型，仅利用2个数据集间的距离差异来识别成员关系。

1.2　威胁模型

成员推理攻击的威胁模型如表1所示，其中，√和×分别指需要或不需要该知识实施攻击。由表1可知，黑盒、灰盒和白盒攻击的背景知识依次增强，攻击者的攻击能力也逐步增强。

表1　　成员推理攻击的威胁模型

威胁模型	训练数据分布	目标模型		
		输入	输出	架构和参数
黑盒	×	√	√	×
灰盒	√	√	√	×
白盒	√	√	√	√

1.3　成员推理防御

已有许多防御机制相继被提出以抵御MIA，如正则化技术[3,8]、差分隐私[9]、信任分数掩蔽技

术[10]、知识蒸馏[11-12]以及成员样本预先排除技术[13]。

1.4　攻击方法概述

表 2 展示了本文研究的 15 种 MIA 方法，其中，√和×分别指需要或不需要该知识实施攻击。BlindMI-w/o[7]和 BlindMI-w 攻击[7]都基于 2 个数据集之间的距离差异性来识别成员关系，主要区别是非成员数据集不同。GradDiff 攻击[14]在联邦学习场景下，利用数据集间的距离差异性来实施攻击。NN 攻击[3]利用多个影子模型的输出训练攻击模型。ImproveMI 攻击[15]利用数据集不同子组的模型输出分别训练多个攻击模型。EnhancedMI 攻击[16]在联邦学习场景下提出了一种增强攻击，与 NN 攻击类似。Top3 NN 攻击[5]利用 Top3 输出概率来实施攻击。Top1 Threshold 攻击[5]基于 Top1 信任分数值来鉴别成员。Loss Threshold 攻击[6]比较目标样本的交叉熵损失与成员样本的平均交叉熵损失的相对大小。Label-Only 攻击[6]采用了预测标签和真实标签的一致性。Top2+True 攻击[7]结合了 Top2 输出概率和真实标签。LiRA 攻击[17]联合了样本的难度校准分数和高斯似然估计来识别成员关系。Risk score 攻击[18]利用样本的隐私风险分数和预测熵来实施攻击。Calibrated score 攻击[19]根据样本的难度校准分数来进行识别。Distillation based 攻击[20]利用蒸馏技术来设置同时依赖模型和数据的攻击阈值。

表 2　　成员推理攻击方法描述

攻击方法	威胁模型			影子模型	攻击模型
	黑盒	灰盒	白盒		
NN[3]	√	×	×	多个	多个
Top3 NN[5]	√	×	×	一个	一个
BlindMI-w[7]	√	√	√	×	×
BlindMI-w/o[7]	√	√	√	×	×
Label-Only[6]	√	×	×	×	×
Top2+True[7]	√	×	×	多个	多个
Top1 Threshold[5]	√	×	×	×	×
Loss Threshold[6]	√	×	×	一个	×
EnhancedMI[16]	√	×	×	多个	多个
GradDiff[14]	√	√	√	×	×
ImproveMI[15]	√	×	×	多个	多个
LiRA[17]	√	×	×	多个	×
Risk score[18]	√	×	×	多个	×
Calibrated score[19]	√	×	×	多个	×
Distillation based[20]	√	×	×	多个	×

2　基于数据特征的成员推理增强攻击方法

针对现有 MIA 并未充分检测攻击有效性，导致错误推测并造成经济损失的问题，本文首先提出了一种成员推理攻击有效性检测算法。然后，根据目标数据、数据特征和攻击原理，提出了一种基于数据特征的成员推理增强攻击方法（如图 1 所示）。

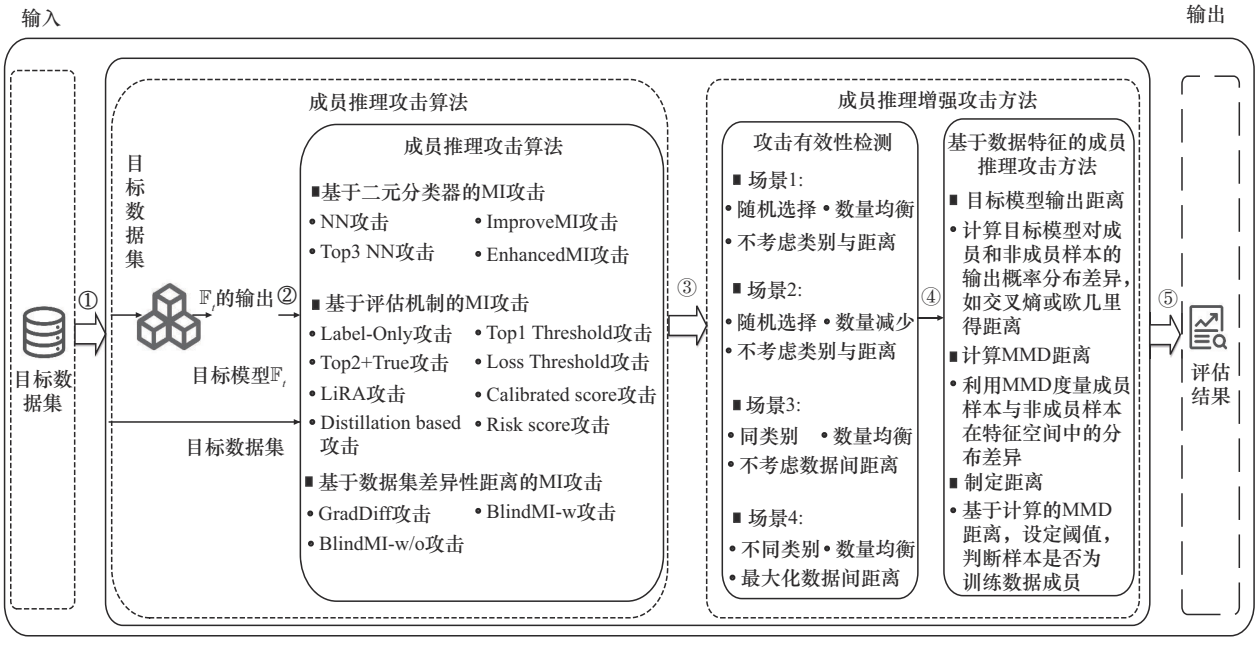


图 1　基于数据特征的成员推理增强攻击方法

2.1 成员推理攻击有效性检测算法设计

1) 对数据样本进行分类

为了检测现有MIA的有效性，本文首先将数据集中的样本分为8类：正确分类的样本、错误分类的样本、成员样本、非成员样本、正确分类的成员样本、正确分类的非成员样本、错误分类的成员样本和错误分类的非成员样本。然后计算这8类样本分别对应的数量，以及攻击的精确率（precision）、召回率（recall，也称真阳率（TPR, true positive rate））、F1-score、假阳率和成员增益。最后，通过综合评估来判断MIA的有效性。

2) 计算数据样本间的距离

给定任意2个数据样本p和q，首先将其输入目标模型获得输出概率。然后将所有输出概率映射到再生核希尔伯特空间[21]，并计算数据样本间的最大均值差异性（MMD, maximum mean discrepancy）距离。

$$\mathrm{Dis}\left[P_p,P_q\right]=\left|\frac{1}{n_p}\sum_{i=1}^{n_p}\phi\left(P_i\right)-\frac{1}{n_p}\sum_{j=1}^{n_q}\phi\left(P_j\right)\right|_v \tag{1}$$

其中，$\phi(P_i)$和$\phi(P_j)$是再生核希尔伯特空间中的描述，v是核空间的维度，$P_i\in P_p$且$P_j\in P_q$，$n_p(n_q)$是样本$p(q)$的模型输出概率$P_p(P_q)$的大小。$\phi(\cdot)$是从概率空间到再生核希尔伯特空间的特征空间映射，即$k\mapsto v$，高斯核函数最常用。

3) 成员推理攻击有效性检测算法

在对数据样本分类和计算距离后，本文构造了4种测试场景，并提出了一种成员推理攻击有效性检测算法。场景1（TS_1）的目标数据集D_{t1}分别在目标训练集和测试集（或生成的非成员数据集）中随机选择相等数量的成员和非成员（如10 000），且不考虑数据集类别和数据样本间的距离。场景2（TS_2）的目标数据集D_{t2}分别在这2个数据集中随机选择相等数量且小于场景1数量的成员和非成员（如6 000），且不考虑数据集类别和数据样本间的距离。场景3（TS_3）的目标数据集D_{t3}分别在这2个数据集中选择相等数量且为同一类别的成员和非成员（如6 000），且考虑数据集类别，但不考虑数据样本间的距离。场景4（TS_4）的目标数据集D_{t4}分别在这2个数据集中选择相等数量且不同类别的成员和非成员（如6 000），并确保数据样本间距离最大。

算法1描述了成员推理攻击有效性检测算法。对于给定的ML模型、数据集和MIA，首先将数据集的数据样本分为8类，计算每类的样本数量以及数据样本间的距离。然后构造4个测试场景，并检测已有MIA的有效性。算法1设置了一个假阳率阈值α和一个成员增益阈值ϖ，即若计算出的假阳率和成员增益满足阈值要求，则认为该MIA有效。设置原理是现实中不同攻击者对假阳率的容忍度和成员增益的要求不同。因此，本文将攻击者能容忍的最大的假阳率作为假阳率阈值α，将攻击者能接受的最小成员增益作为成员增益阈值ϖ。

2.2 基于数据特征的成员推理增强攻击方法设计

利用成员推理攻击有效性检测算法检测现有MIA的有效性后，为了降低攻击的假阳率并提高攻击的有效性，本文根据攻击时使用的目标数据、数据特征和攻击原理，提出了一种基于数据特征的成员推理增强攻击方法。数据特征主要包括目标模型对数据的输出概率、数据样本间的距离（如MMD距离）。这些数据特征是数据自身的特征，并没有涉及对数据处理后提取的特征。该方法的原理是基于模型对数据的输出概率以及输出概率之间的关联性（如MMD距离），来挖掘数据遭受攻击的风险，进而选择合适的数据来增强攻击效果。

算法1 成员推理攻击有效性检测

定义$i=1$，给定ML模型，不同数据集D_{ti}和成员推理攻击MIA_j，假阳率阈值α，成员增益阈值ϖ

1) 对于第j个成员推理攻击MIA_j
2) 循环
3) 将数据集D_{ti}输入目标ML模型
4) 将数据集D_{ti}分为8类并计算每类的样本数量
5) 构造测试场景$\mathrm{TS}=\left\{\mathrm{TS}_1,\mathrm{TS}_2,\mathrm{TS}_3,\mathrm{TS}_4\right\}$
6) for $\mathrm{TS}_i\in\mathrm{TS}$
7) 计算precision、recall、F1-score、MA和FPR
8) $i=i+1$
9) 循环
10) until $i=4$，保存计算的5个评估指标数值
11) 循环
12) until j等于成员推理攻击的数量，保存计算的评估指标的数值
13) 如果MA>ϖ且FPR < α
14) 返回攻击方法有效
15) 否则
16) 返回攻击方法无效
17) end for

算法 2　基于数据特征的成员推理增强攻击

定义 $i = 1$，给定 ML 模型，不同的目标数据集 D_{ti} 和成员推理攻击 MIA_j，最小假阳率 δ，初始化可测试的目标数据集 D_{test}

1) 对于不同的目标数据集 D_{ti}

2) 循环

3) 将数据集 D_{ti} 输入目标 ML 模型

4) 根据 ML 模型的输出将数据集 D_{ti} 分为两类

5) 用式 (3) 计算数据样本间的 MMD 距离 MMD_{hl}

6) 对于第 j 个成员推理攻击 MIA_j

7) 从算法 1 获得假阳率为 δ 对应的 MMD 距离，记作距离阈值 π

8) 如果 MMD_{hl} 大于 π，即 $\mathrm{MMD}_{hl} > \pi$ 成立

9) 则算法收敛，将此时的目标数据集存入可测试的目标数据集，即 $D_{\text{test}} \leftarrow \{D_{ti}\}$

10) 否则，算法未收敛，利用数据生成技术生成一些非成员数据

11) 将生成的非成员数据与“输出概率低的样本”进行替换，重新组成一个新的测试数据集 D_{ti}^{n}

12) 用式(3)计算 MMD_{hn}

13) 如果 MMD_{hn} 大于 π，即 $\mathrm{MMD}_{hn} > \pi$ 成立

14) 则算法收敛，$D_{\text{test}} \leftarrow \{D_{ti}^{n}\}$

15) 返回可测试的目标数据集 D_{test}

算法2描述了基于数据特征的成员推理增强攻击方法，对于一个目标数据集，首先根据目标模型对目标数据的输出概率将数据分为两类——“输出概率高的样本”和“输出概率低的样本”。接着用式(1)计算这两类数据样本间的 MMD 距离 MMD_{hl}，并根据攻击者能容忍的最小假阳率 δ，利用算法 1 计算该最小假阳率下数据样本间的 MMD 距离，将其作为距离阈值 π，若 MMD_{hl} 大于 π，则说明假阳率小于 δ，即该目标数据集可用。否则，利用数据生成技术[7]生成一些非成员数据，并将该非成员数据和“输出概率低的样本”进行替换，然后再计算“输出概率高的样本”和生成的非成员数据 MMD 距离 MMD_{hn}，并判断 MMD_{hn} 与距离阈值 π 的相对大小，该操作一直进行到 MMD_{hn} 大于 π。由于目标数据集中数据样本的成员关系无法事先预知，本文首先根据模型输出概率将数据样本分为上述两类，同时已有研究[3,5]表明目标模型对成员样本的表现优于非成员样本，因此本文认为“输出概率高的样本”更有可能是成员，“输出概率低的样本”更有可能是非成员。在对数据样本进行替换时，首先将“输出概率高的样本”的索引固定住，只将生成样本与“输出概率低的样本”进行替换，并确保替换后数据样本与“输出概率高的样本”间的距离大于距离阈值 π。最终实现攻击者可接受的最小假阳率下的攻击效果。

3　实验评估

本节首先介绍了实验设置、模型架构、评估指标以及数据集，然后评估了本文所提鲁棒性检测方法和攻击增强方法的有效性。

3.1　实验设置

本文实验的操作系统是 Ubuntu 22 64 位，CPU 是 AMD EPYC 7542 32-Core Processor，GPU 是 NVIDIA GeForce RTX 3090，CUDA 版本是 12.0，Python 版本是 3.10，Pytorch 版本是 2.3.0。

3.2　模型架构和评估指标

如表 3 所示，参考 BlindMI 攻击[7]本文实验采用了多层感知器（MLP, multilayer perceptron）模型，

表 3　目标模型和影子模型架构以及超参数设置

模型架构	层数	目标模型		影子模型	
		最大的 Epoch	学习率	最大的 Epoch	学习率
MLP	3~7	e_m	5×10^{-5}	$[0.3\sim2]e_m$	1×10^{-4} 或 1×10^{-5}
StandDNN	2	e_m	5×10^{-5}	$0.5e_m$	1×10^{-4}
VGG16	16	$e_p + e_m$	5×10^{-5}	$e_p + 0.6e_m$	5×10^{-5}
VGG19	19	$e_p + e_m$	5×10^{-5}	$e_p + 1.5e_m$	5×10^{-5}
ResNet50	50	$e_p + e_m$	5×10^{-5}	$e_p + 0.2e_m$	5×10^{-5}
ResNet101	101	$e_p + e_m$	5×10^{-5}	$e_p + 0.3e_m$	1×10^{-4}
DenseNet121	121	$e_p + e_m$	5×10^{-5}	$e_p + e_m$	1×10^{-4}

最多有7个密集层（8 192、4 096、2 048、1 024、512、256和128）和一个Softmax层，其中，e_m表示目标模型在表4的每个数据集上的最大Epoch，e_p表示在ImageNet数据集上预训练权重的Epoch。标准的卷积神经网络（CNN, convolutional neural network）架构和超参数与已有工作[3]相同，还利用了3种流行的深度神经网络（DNN, deep neural network）模型。给定一个训练集，实验随机选择一个目标模型，并用指定超参数训练该目标模型。黑盒攻击随机选择一个影子模型架构，灰盒和白盒攻击需选择与给定目标模型相同的影子模型。

3.3　数据集

如表4所示，参考已有工作[7,18]，本文实验采用MIA领域4个广泛使用数据集：CIFAR100、CIFAR10、CH_MNIST和ImageNet，且测试集中成员和非成员数量相等。

3.4　攻击有效性检测算法的有效性

本文在4个测试场景上测试各个攻击的效果，当测试的假阳率小于α且成员增益大于ϖ时，将其视为攻击有效，反之亦然。本文实验设置α和ϖ的取值范围分别是{5%，10%，15%，20%}和{20%，30%，40%，50%}。从图2~图9可知，尽管已有

表4　　**实验中使用的数据集**

数据集	场景	类别	分辨率	Epoch大小	训练数据集	参考数据集	测试集成员	测试集非成员
CIFAR10	图像分类	10	32×32	(pre-trained + 30)或150	25 000	25 000	12 500	12 500
CIFAR100	图像分类	100	32×32	(pre-trained + 30)或150	25 000	25 000	12 500	12 500
CH_MNIST	图像分类	8	64×64	(pre-trained + 15)或150	2 250	2 250	500	500
ImageNet	图像分类	200	64×64	(pre-trained + 30)或150	10 000	10 000	20 000	20 000

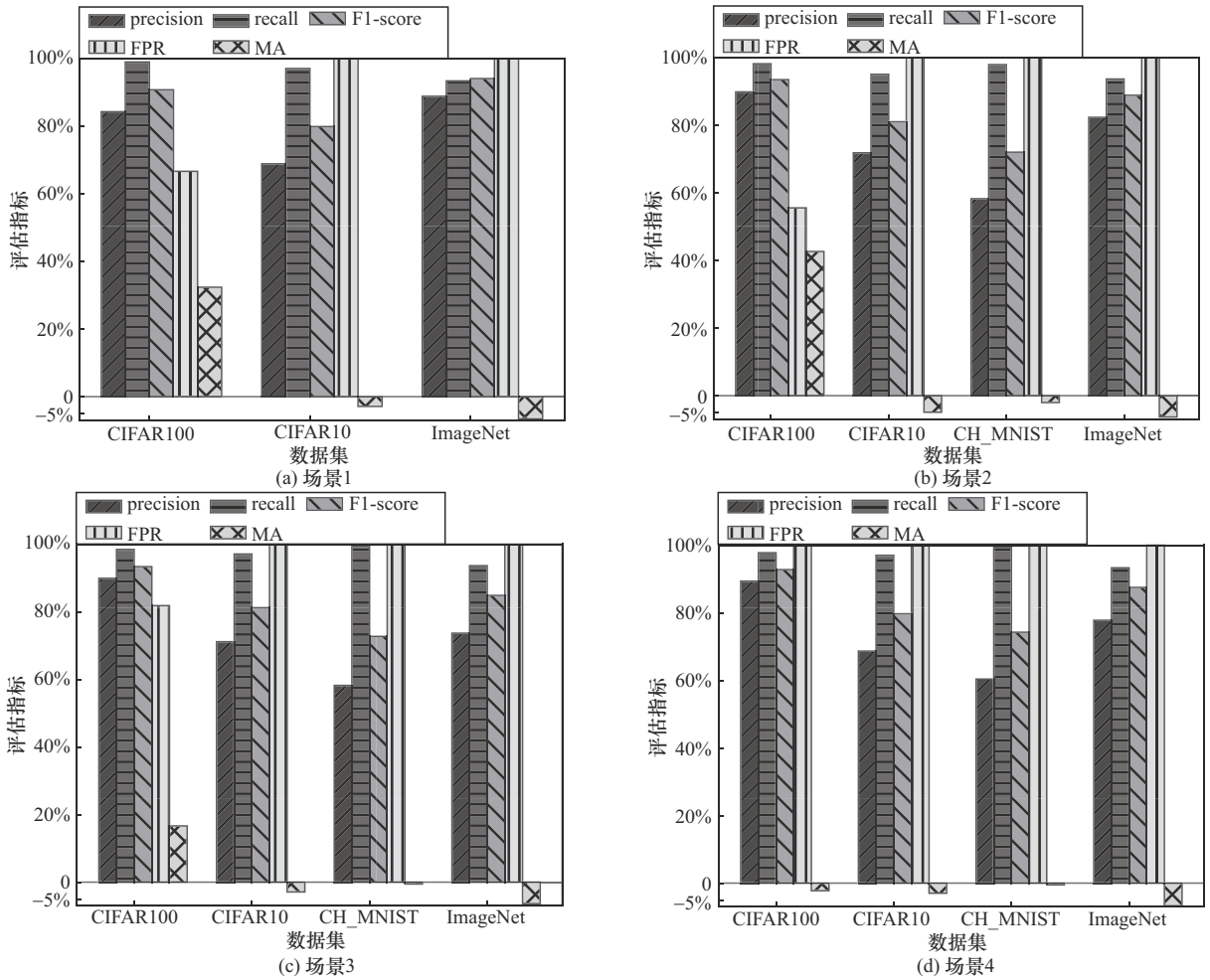


图2　BlindMI-w/o攻击在4种测试场景中的攻击效果

MIA 在不同测试场景下的 F1-score（如 87.49%）和召回率（如 99.17%）较高，但是其假阳率（如 100%）却远远超出了假阳率阈值 α，进一步限制了其实际应用。例如，图 2(a)~图 2(d) 分别展示了 BlindMI-w/o 攻击在场景 1、场景 2、场景 3 和场景 4 中的攻击效果，可以看出虽然 BlindMI-w/o 攻击的召回率和 F1-score 较高（如 99.04%），但其假阳率最高可达 100%，而攻击者能接受的假阳率通常为 10% 或者更低。因此，BlindMI-w/o 攻击的实际应用不是很理想。图 3 表明，尽管 BlindMI-w 攻击在 CIFAR10 和 CH_MNIST 数据集上的召回率和 F1-score 较高（如 92.77%），但其仍存在较高的假阳率（如 75%）。GradDiff 攻击也基于数据集差异来实施攻击，表现出了与 BlindMI-w 攻击相似的攻击效果。

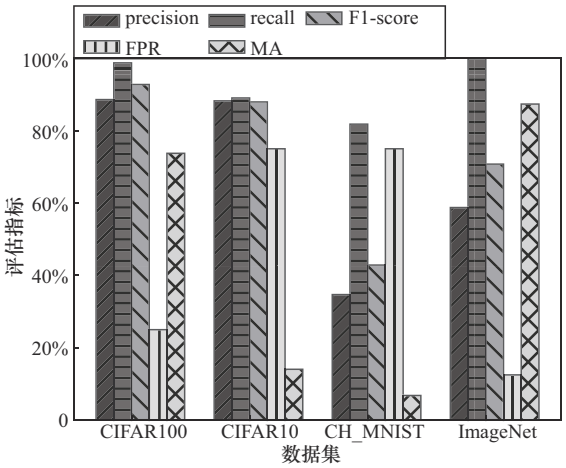


图 3　BlindMI-w 攻击在测试场景 1 下的攻击效果

由图 4 可知，当测试数据集的数据量发生变化时，Label-Only 攻击仍表现出较高的假阳率。例如，当图 4(a) 中测试数据为 10 000 时，Label-Only 攻击在 CH_MNIST 上的假阳率达到了 81.04%。当图 4(b) 中测试数据为 6 000 时，Label-Only 攻击在 CIAFR10 和 CIFAR100 数据集上的假阳率都达到了 100%，而实际中这么高的假阳率是不可用的。图 5 表明，尽管 Loss Threshold 攻击的召回率（如 99.07%）和 F1-score（如 87.35%）较高，但是其精确率（如 51.46%）和成员增益（如 9.95%）较低。经分析发现 Loss Threshold 攻击的假阳率较高，如在测试场景 1 和场景 2 上，Loss Threshold 攻击在 ImageNet 数据集上的假阳率分别达到了 94.77% 和 100%，在实际中是不可接受的。观察图 6 可知，NN 攻击在 ImageNet 数据集上召回率达到了 99.42%，但其精确率和成员增益较低（分别为 50.05% 和 −3.58%），究其原因发现其假阳率达到了 100%，极大地损害了攻击的有效性。同时可发现当测试数据集的大小发生变化时，对其攻击有效性的影响较小。EnhancedMI 和 ImproveMI 攻击的攻击效果与 NN 攻击基本一致。

图 7 表明，Top1 Threshold 攻击的 F1-score 最高可达 79.59%，但其成员增益却只有 18.63%，而假阳率高达 93.33%，制约其实际可行性。由图 8 可知，只关注 Top2+True 攻击的召回率和 F1-score 时（如 98.36% 和 70.79%），可认为其是有效的，但其精确率和成员增益分别为 55.30% 和 18.84%。因此实际中 Top2+True 攻击并非有效。图 9 表明 Top3 NN 攻击的 F1-score 和召回率较高（91.12%），而其成员增益较低（26.24%），原因在于不同测试场景下该攻击的假阳率较高（如 100%），超出了承受范

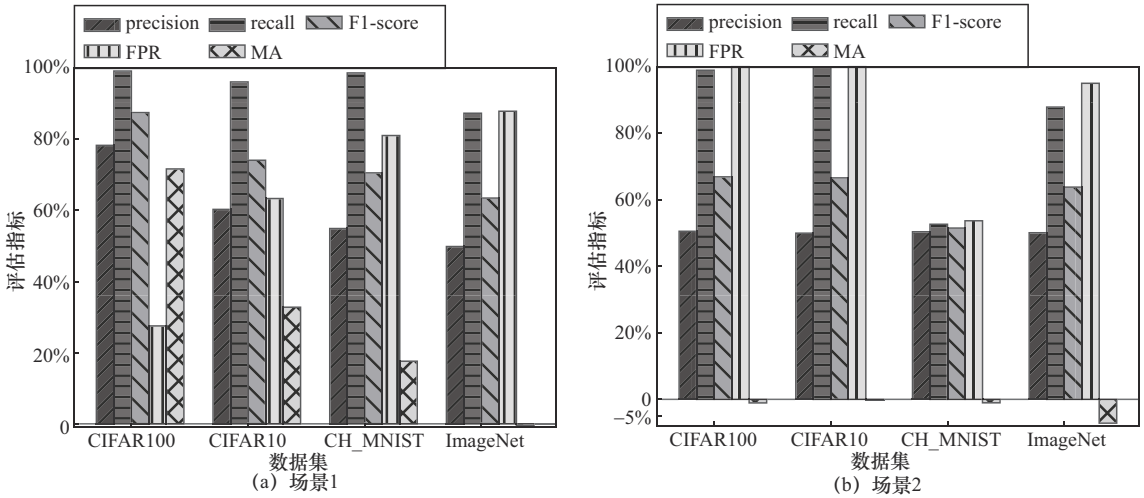
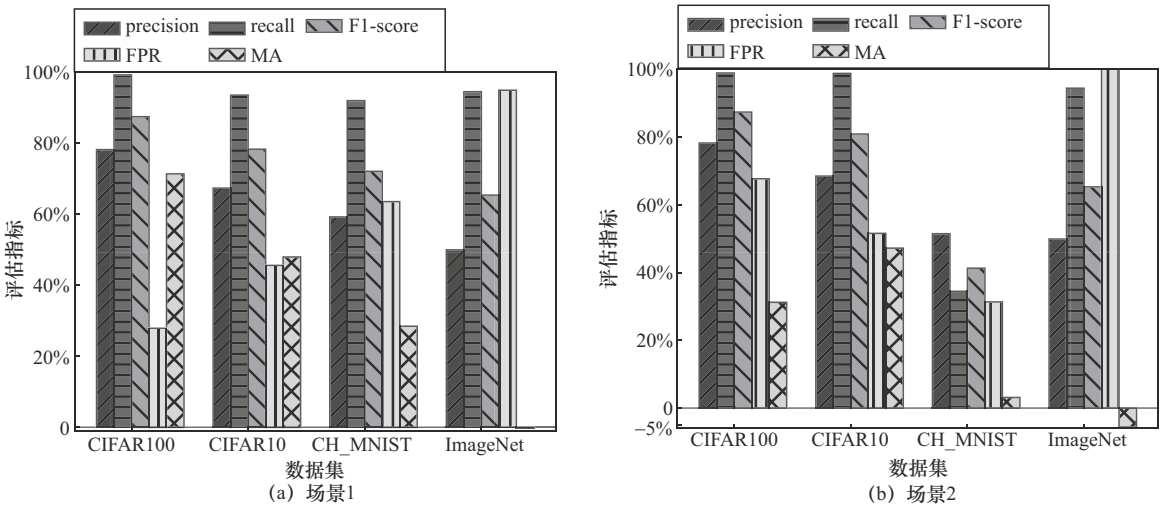
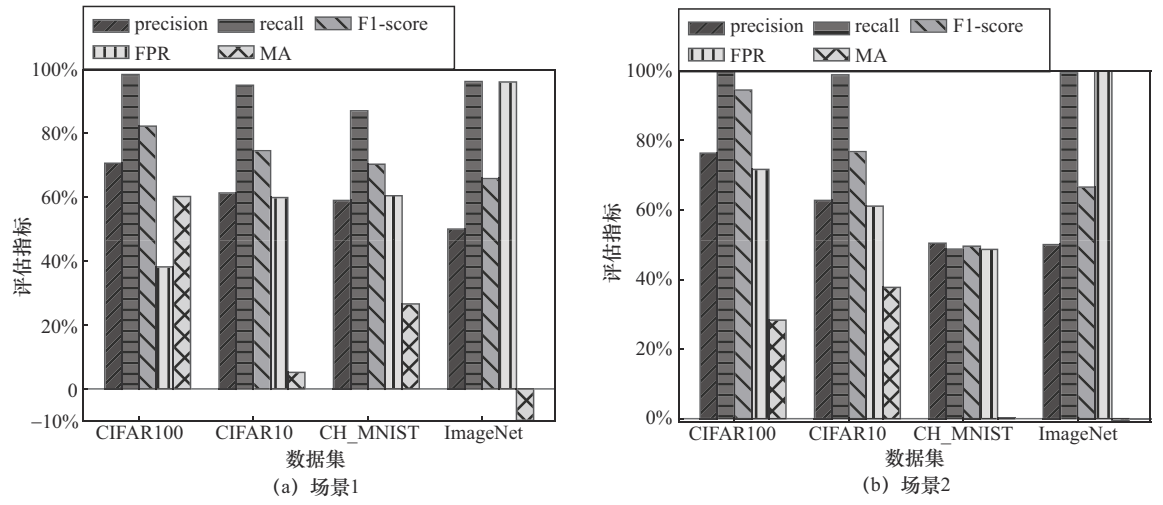


图 4　Label-Only 攻击在测试场景 1 和场景 2 中的攻击效果



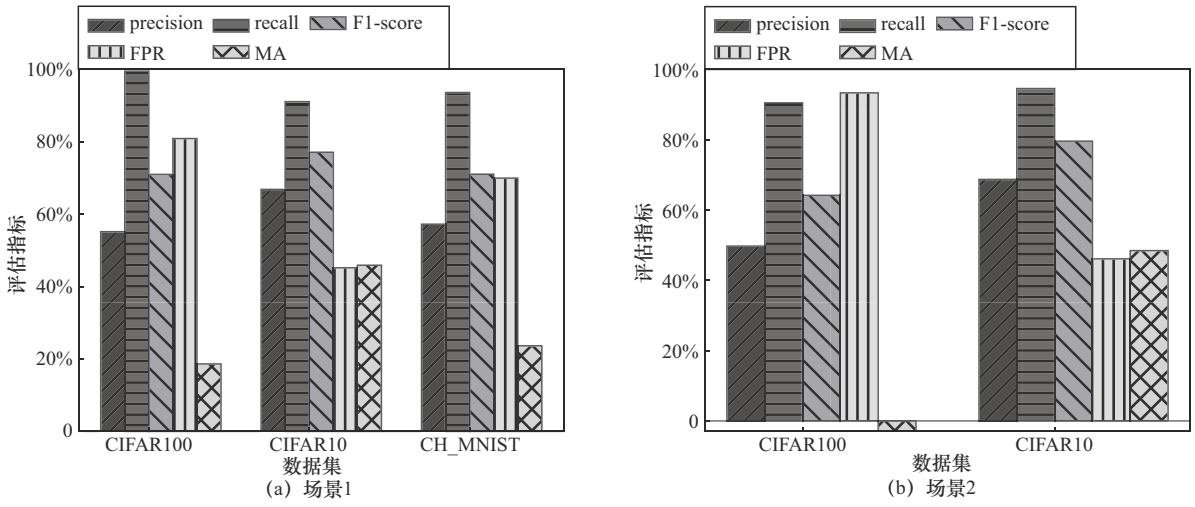
(a) 场景1　　(b) 场景2

图5　Loss Threshold攻击在测试场景1和场景2中的攻击效果



(a) 场景1　　(b) 场景2

图6　NN攻击在测试场景1和场景2中的攻击效果



(a) 场景1　　(b) 场景2

图7　Top1 Threshold攻击在测试场景1和场景2中的攻击效果

围。同时，观察表5可知，不同攻击的曲线下面积（AUC, area under the curve）和TPR@0.1%FPR这2个指标的取值均较低，也从侧面揭示了已有工作有较高的假阳率。

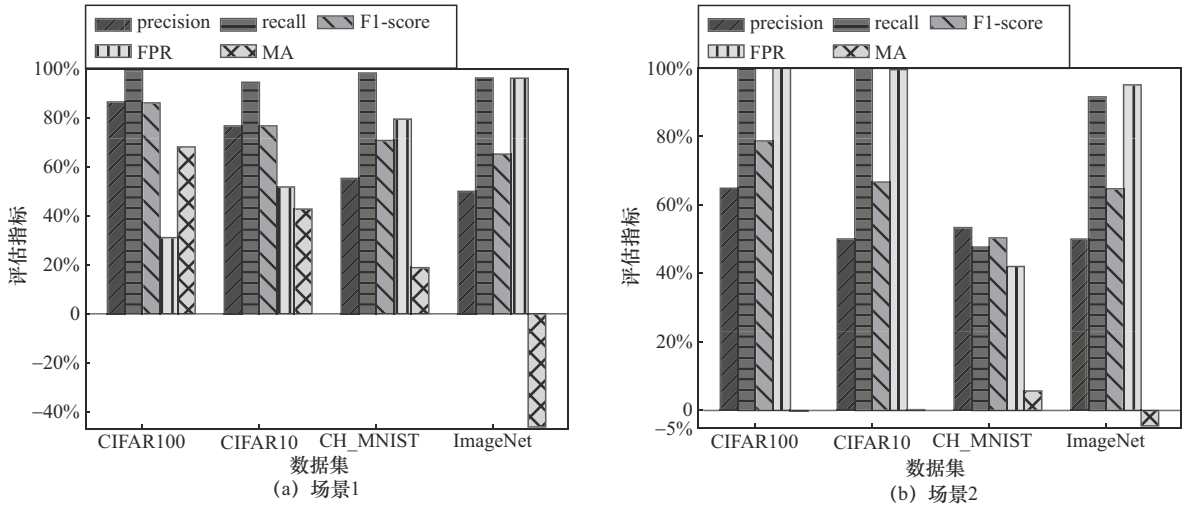


图 8　Top2+True 攻击在测试场景 1 和场景 2 中的攻击效果

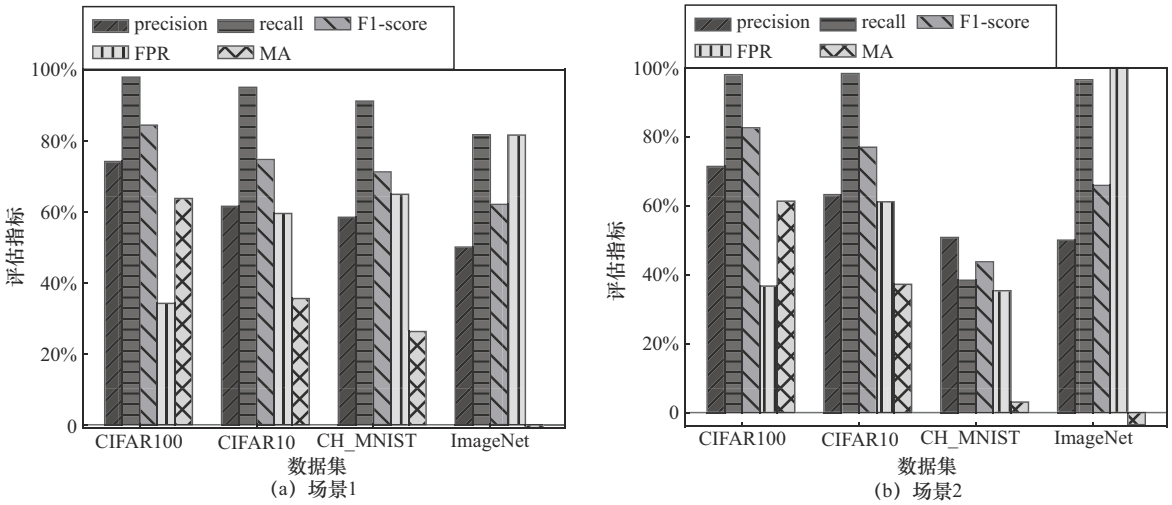


图 9　Top3 NN 攻击在测试场景 1 和场景 2 中的攻击效果

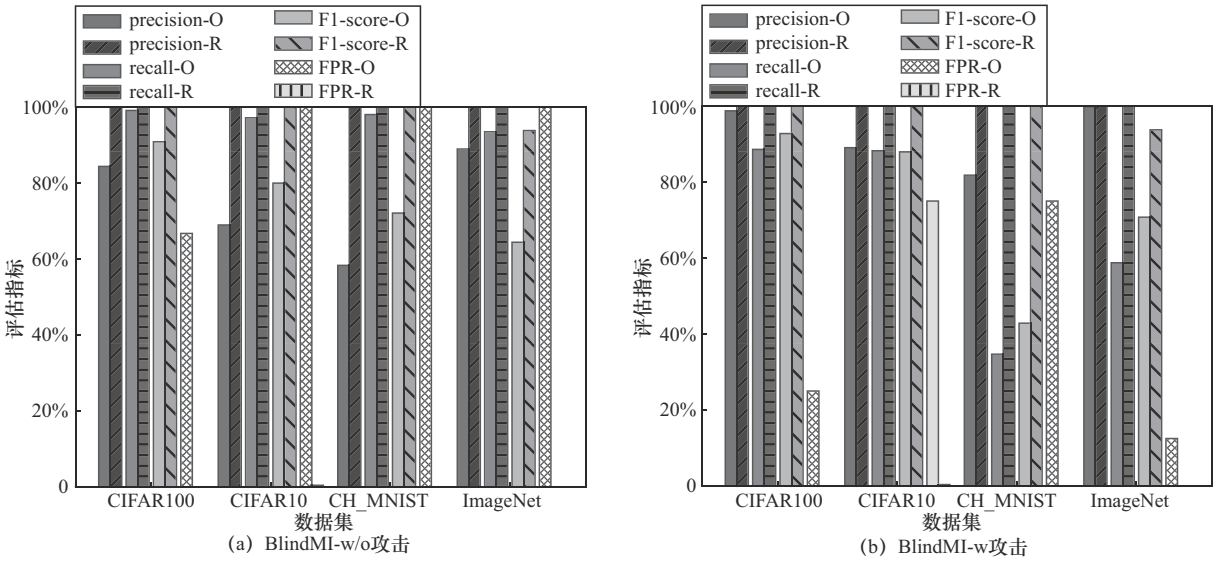
表5　不同攻击方法AUC和TPR@0.1%FPR的测试结果

攻击方法	AUC	TPR@0.1%FPR
NN	0.53	0.12%
Top3 NN	0.52	0
Label-Only	0.55	0
Top1 Threshold	0.828	0
BlindMI-w	0.501	0
Top2+True	0.597	0

3.5　基于数据特征的成员推理增强攻击的有效性

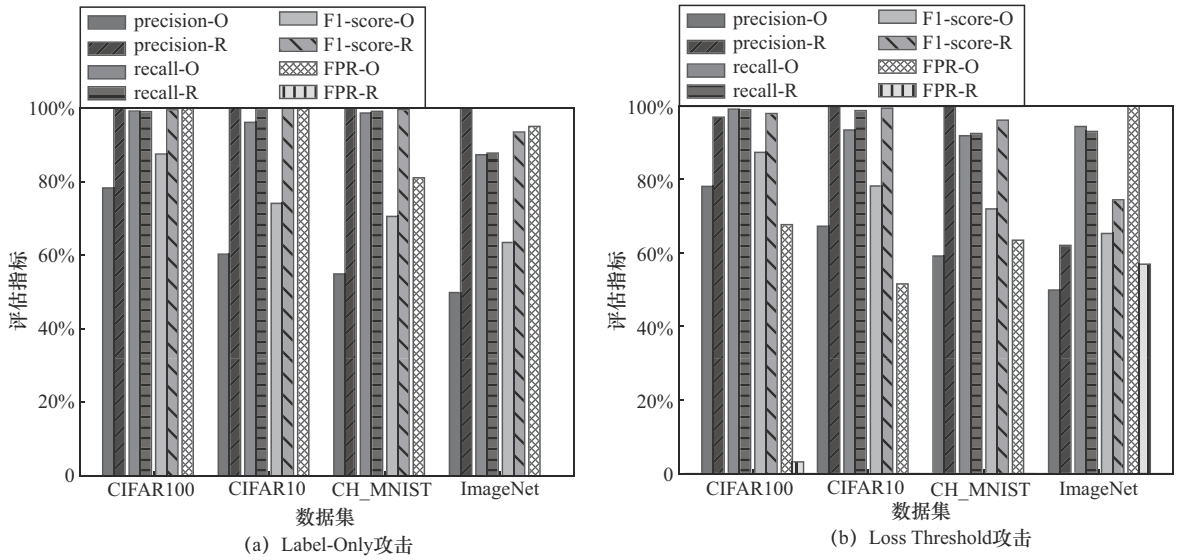
本节主要在 4 个数据集上分别对比了利用本文所提的增强攻击方法处理前（O, original）和处理后（R, reduction），15 种 MIA 的攻击效果。图 10 表明，本文方法在降低 BlindMI-w/o 和 BlindMI-w 攻击假阳率的同时，将这 2 个攻击的成员增益分别提升了 67.63% 和 36.4%。由此可见，本文方法对 BlindMI-w/o 攻击的性能提升力度更大，主要原因是 BlindMI-w/o 攻击直接利用测试数据集作为非成员数据集来捕捉成员关系，而 BlindMI-w 攻击则利用生成的非成员数据集。而生成的非成员数据与替换数据都是生成数据（可能采用相同的数据生成技术），导致替换后数据间的距离相较于测试数据与生成数据的有所减小，从而限制了 BlindMI-w 攻击性能提升。同时，GradDiff 与 BlindMI-w 攻击的性能提升幅度基本一致。

图 11 表明在 CIFAR100 数据集上，相较于 Loss Threshold 攻击，本文方法对 Label-Only 攻击的性能增强力度更大。例如，将 Label-Only 攻击精确率和成员增益分别提升了 10.38% 和 55.27%。原因是 Label-Only 攻击利用预测标签与实际标签的一致性来



(a) BlindMI-w/o攻击　　(b) BlindMI-w攻击

图10　BlindMI-w/o和BlindMI-w攻击处理前和处理后攻击效果对比



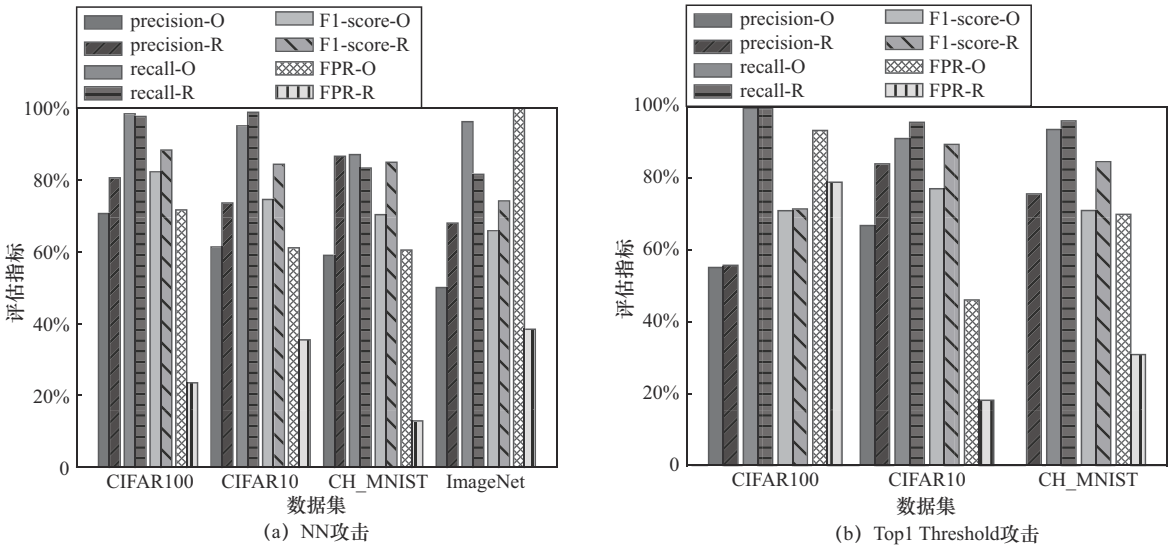
(a) Label-Only攻击　　(b) Loss Threshold攻击

图11　Label-Only和Loss Threshold攻击处理前和处理后攻击效果对比

识别成员，Loss Threshold攻击将目标样本的交叉熵损失与成员样本的平均交叉熵损失的相对大小作为成员依据。在用生成的非成员数据对测试数据中“输出概率低的样本”进行替换时，由于替换后数据标签保持不变，且替换后数据距离增大了，降低了样本间的区分难度，极大增强了Label-Only攻击的性能。而替换的样本虽确保了数据间的差异性，但无法获取成员样本的交叉熵损失，因而对Loss Threshold攻击的性能提升相对较小。

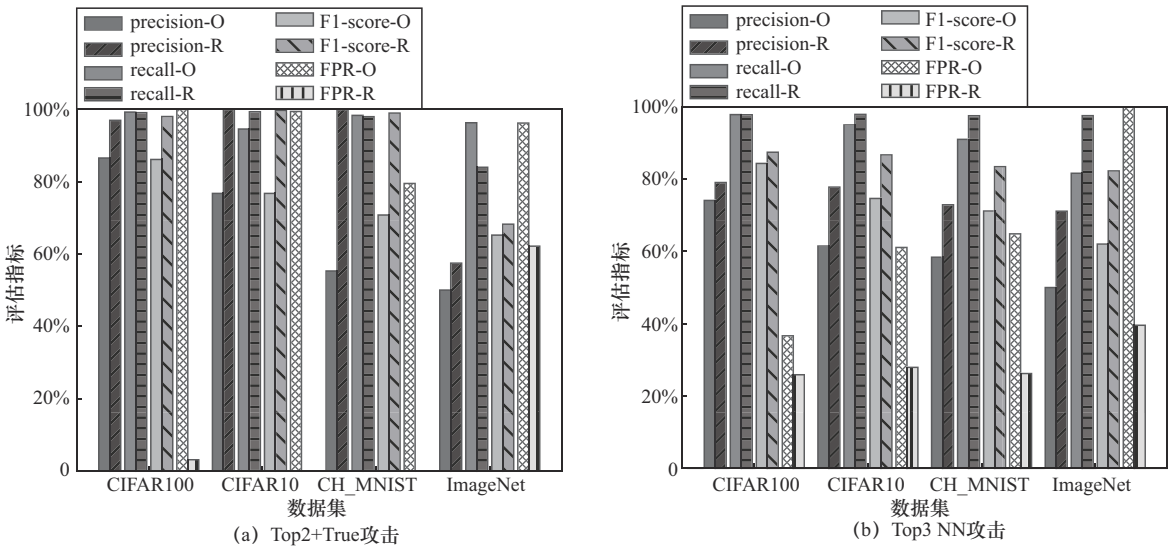
如图12所示，相较于Top1 Threshold攻击，本文方法可将NN攻击的假阳率降低幅度较大，而对这2个攻击的成员增益增强幅度大致相当（如46.91%和41.47%）。主要原因是NN攻击是基于模型的全部输出概率来实施攻击，替换后的样本保证了测试数据间的距离最大，增大了数据样本间的差异，促使该攻击更易捕捉成员关系，进而提升了攻击性能。而替换的样本很难精准刻画Top1非成员阈值，进而对Top1 Threshold攻击的假阳率降低和性能增强幅度相对NN攻击低。由于EnhancedMI和ImproveMI攻击与NN攻击的原理一致，因此这2个攻击增强效果与NN攻击的基本一致。

图13表明，相对于Top3 NN攻击，本文方法对Top2+True攻击的成员增益增强幅度稍大。经分析发现本文方法可对Top2+True攻击的F1-score增



(a) NN攻击　　(b) Top1 Threshold攻击

图 12　NN 和 Top1 Threshold 攻击处理前和处理后攻击效果对比



(a) Top2+True攻击　　(b) Top3 NN攻击

图 13　Top2+True 和 Top3 NN 攻击处理前和处理后攻击效果对比

强幅度（如 28.2%）大于对 Top3 NN 攻击的增强幅度（如 20.25%）。主要原因是替换的样本会改变测试样本间的距离，对这 2 个攻击均有影响。又由于 Top3 NN 攻击仅利用 Top3 输出概率，而 Top2+True 攻击不仅采用了 Top2 输出概率还联合了真实标签，因此本文方法对 Top2+True 攻击的假阳率降低幅度（如 99.44%）相较 Top3 NN 攻击较大（如 60.42%）。此外，Top2 相较于 Top3 输出概率限制较宽松，从而进一步增强了 Top2+True 攻击的性能。

如图 14 所示，相较于 LiRA 攻击，Distillation based 攻击精确率增大幅度提升了 91.31%。主要原因是 Distillation based 攻击同时依赖模型和数据的攻击阈值来进行成员关系识别，而本文方法不仅利用数据生成技术生成测试数据来捕捉数据特征。而且对“输出概率低的样本”进行替换，并确保替换后模型输出的数据样本间距离较大，其进一步捕捉了模型特征，使构建的攻击阈值更精确，进而对 Distillation based 攻击的性能提升较大。而 LiRA 攻击联合了样本的难度校准分数和高斯似然估计来识别成员关系。尽管替换的样本可以在一定程度上提升该攻击性能，但由于原始训练集无法预知，使对 LiRA 攻击性能提升相对较低。

图 15 表明，相较于 Calibrated Score 攻击，Risk score 攻击的 F1-score 增大幅度提升了 58.28%。主要原因是 Risk score 攻击联合样本的隐私风险分数和预测熵来实施攻击。本文方法仅对“输出概率

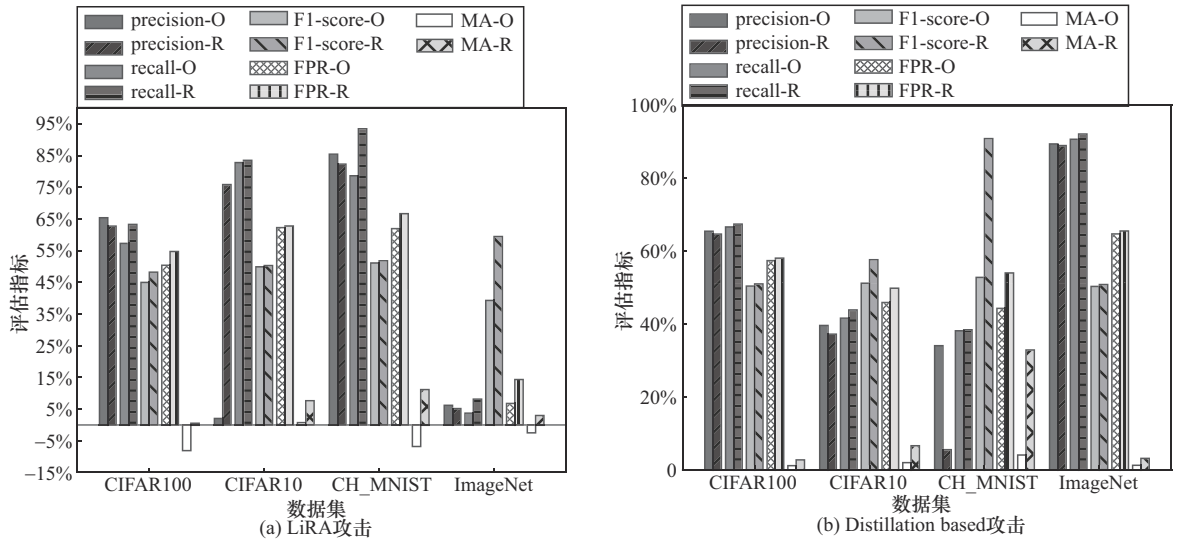


图14　LiRA和Distillation based攻击处理前和处理后攻击效果对比

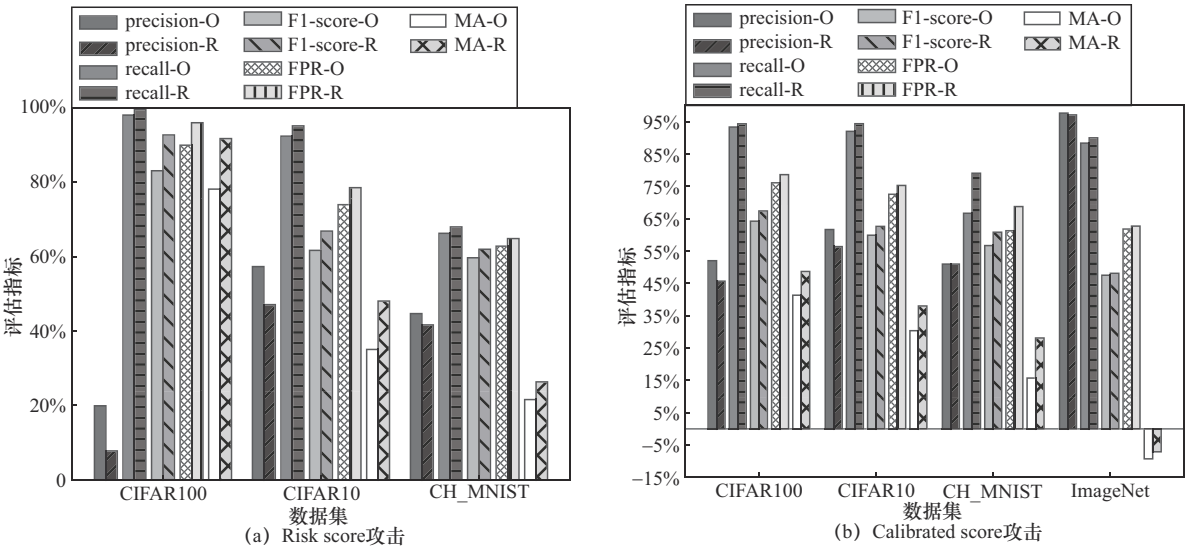


图15　Risk score和Calibrated score攻击处理前和处理后攻击效果对比

低的样本”进行替换，并确保替换后数据样本间的距离较大，增大了测试数据集中隐私风险高的目标样本与低隐私风险样本间的差异，进而增强其攻击性能。而替换后的数据可辅助Calibrated score攻击计算更精确的难度校准分数，又由于训练集未知，因此对Calibrated score攻击的性能提升相对较小。

3.6　消融实验

由图16可知，当数据集需要替换的数据增多时，完成一次攻击所需的时间也有所增加。由于本文仅根据目标模型对数据的输出概率来计算MMD距离，并样本进行替换和攻击。因此，数据替换对攻击整体的执行时间影响不大。图17表明，随着数据替换比例增加，攻击效果逐渐增强。主要原因是替换的数据样本越多，数据样本间的距离越大，攻击越容易捕捉和识别成员和非成员的特征。

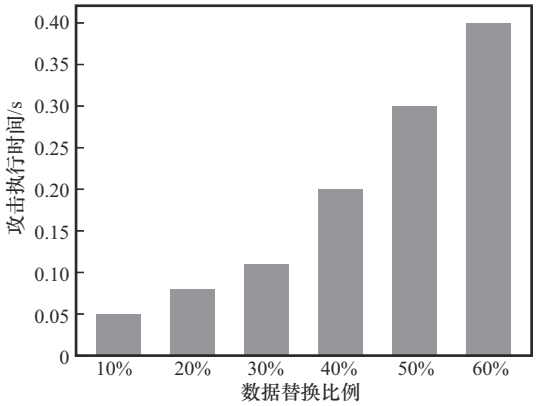
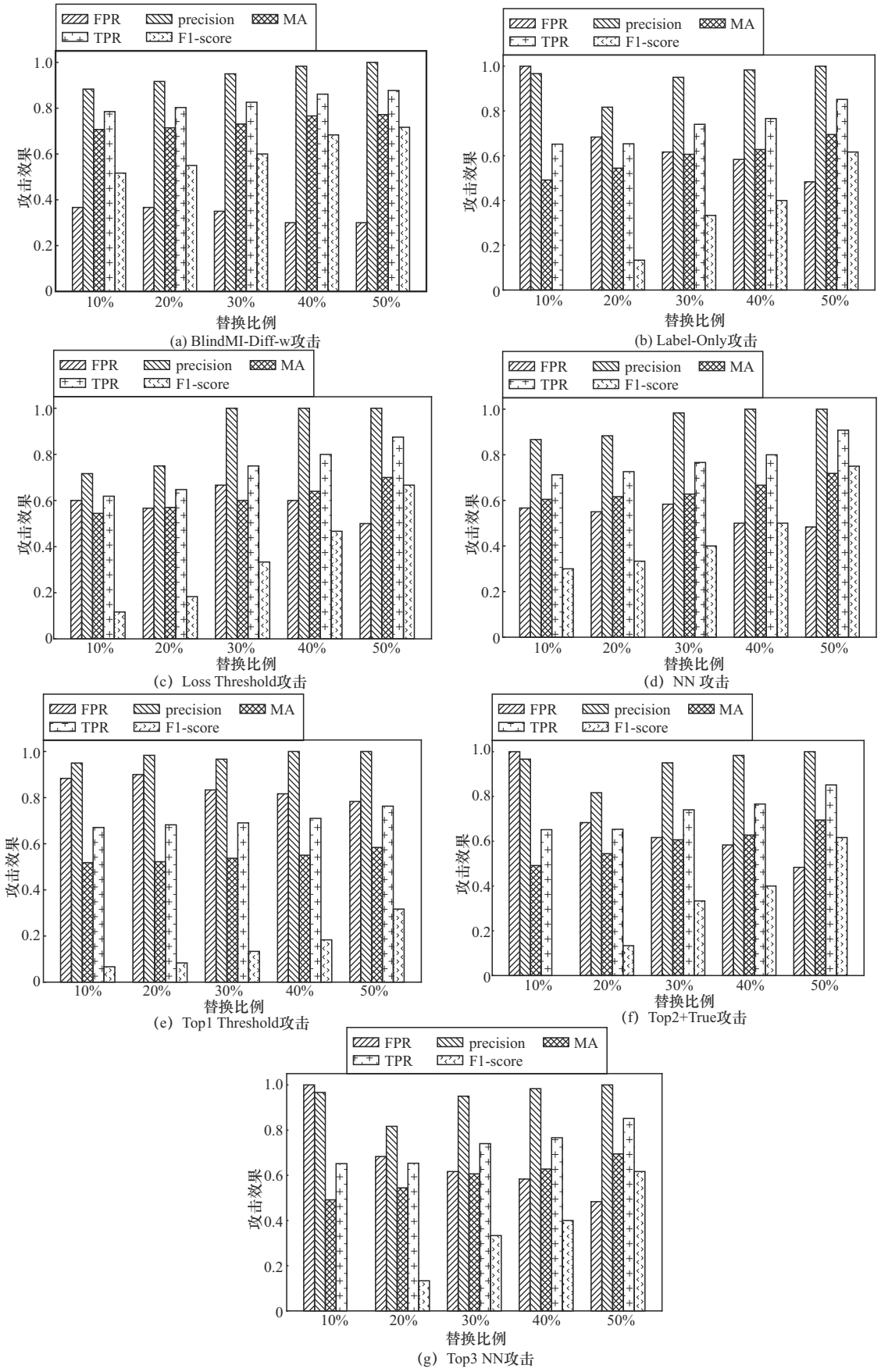


图16　数据替换比例与攻击执行时间的关系



(a) BlindMI-Diff-w攻击　(b) Label-Only攻击

(c) Loss Threshold攻击　(d) NN 攻击

(e) Top1 Threshold攻击　(f) Top2+True攻击

(g) Top3 NN攻击

图17　对数据集进行不同比例的替换后对攻击效果的影响

4　方法讨论

上述研究表明，已有15种MIA仅仅在场景1和场景2下测试攻击效果，并未考虑场景3和场景4的情况，本文在构造的4个测试场景上发现现有15种MIA的假阳率最高可达100%。此外，实验还观察到相比于场景1和场景2，本文所提的有效性检测算法在检测场景3上的假阳率更高。例如，在CIFAR100数据集的场景1和场景3上，BlindMI-w/o攻击的假阳率分别为66.67%和81.82%。图18~图21展示了15种MIA在原始场景（没有任何改进）和改进场景（本文方法）下的攻击效果。

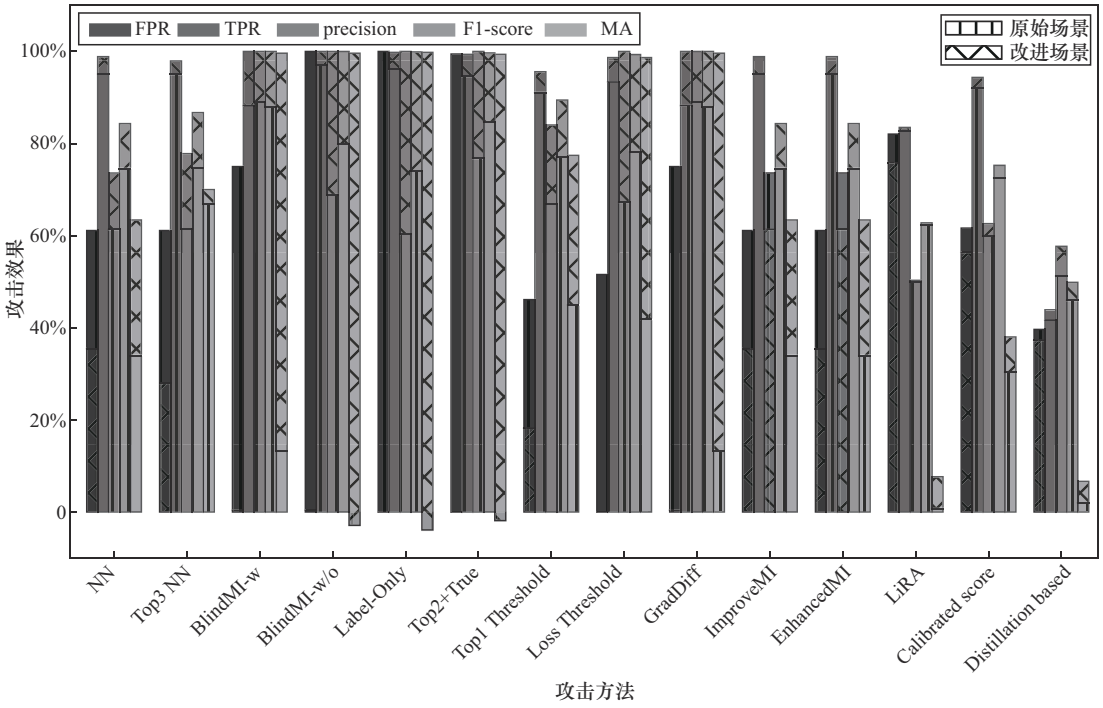


图18　CIFAR100上15种MIA在原始场景和改进场景下的攻击效果

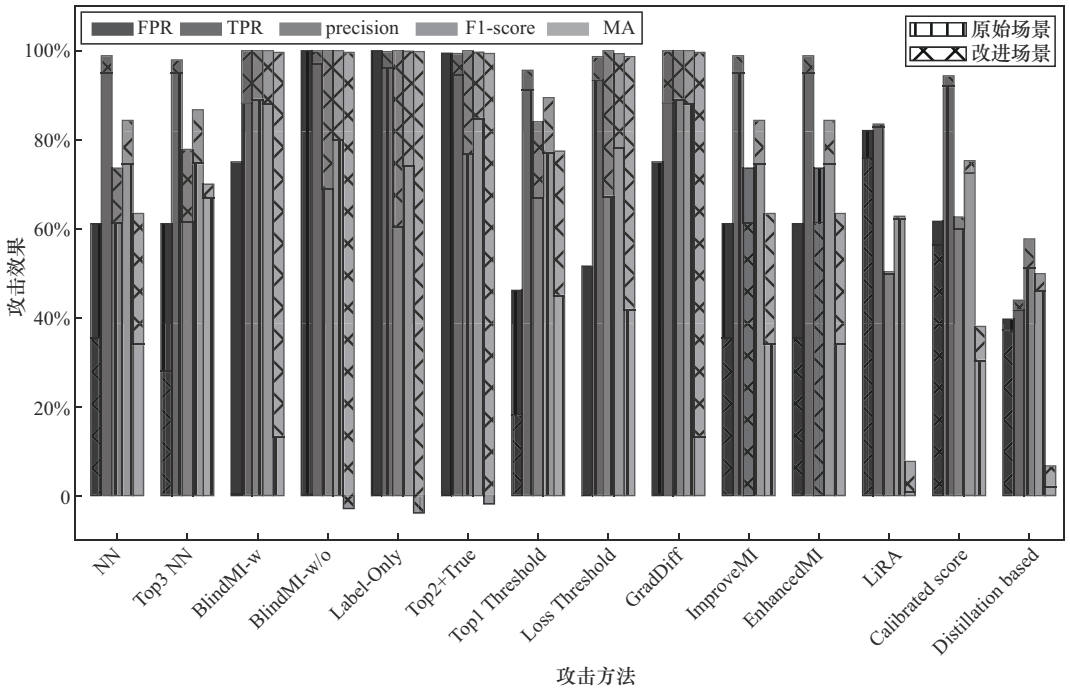


图19　CIFAR10上15种MIA在原始场景和改进场景下的攻击效果

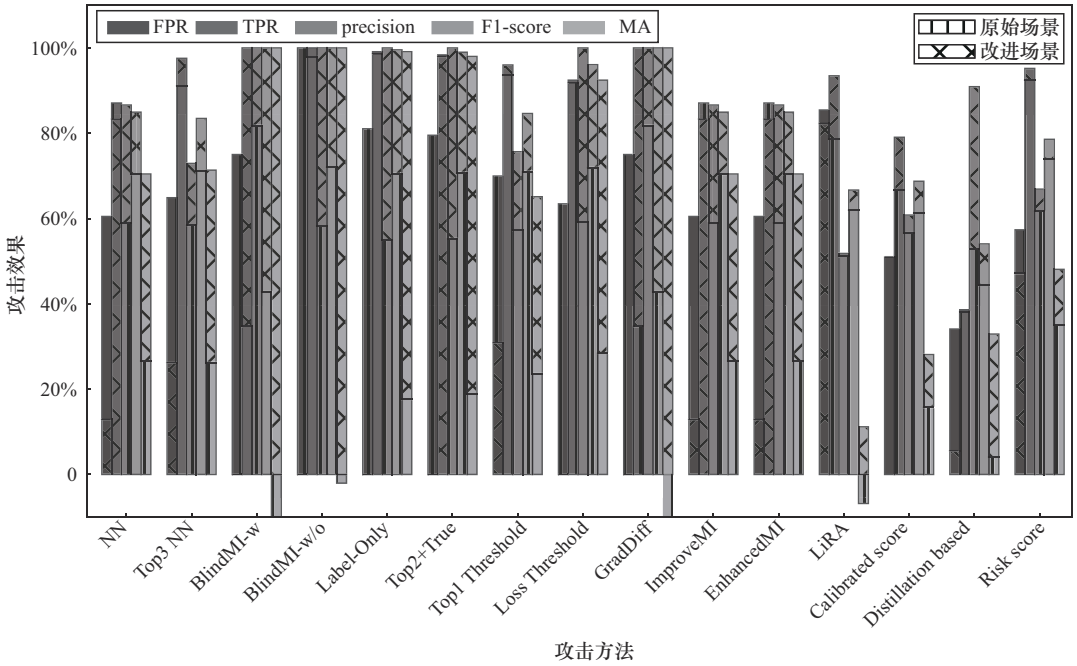


图20　CH_MNIST上15种MIA在原始场景和改进场景下的攻击效果

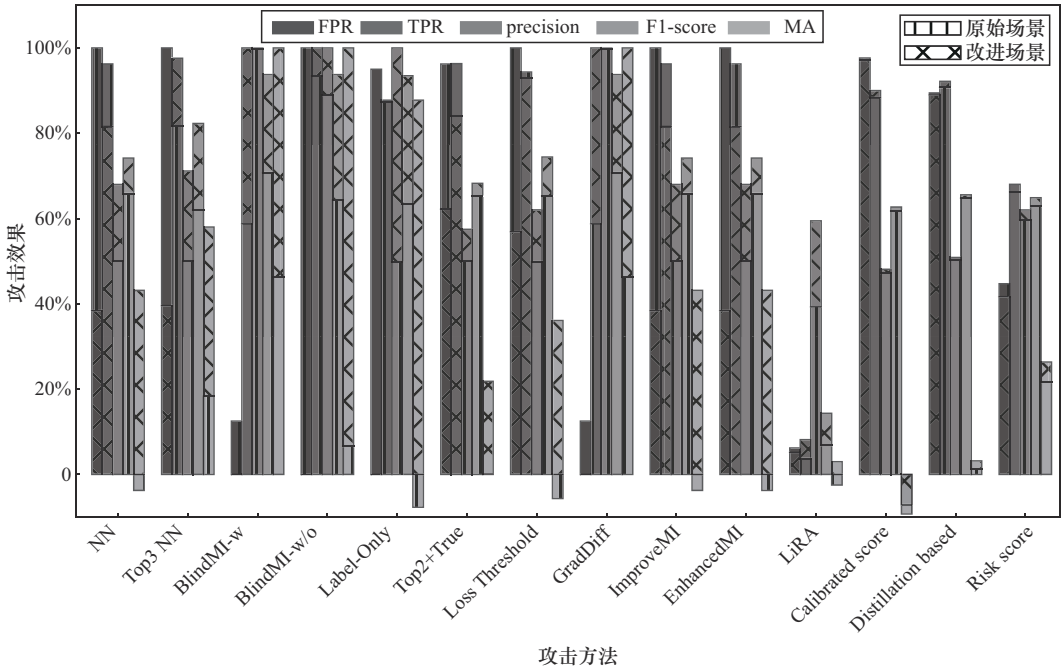


图21　ImageNet上15种MIA在原始场景和改进场景下的攻击效果

由图18~图21可知，相比于原始场景，在利用本文方法改进后的场景上，15种MIA在4个数据集上的攻击效果有显著提升。例如，在CIFAR100数据集上，可将BlindMI-w攻击的精确率、召回率、F1-score、成员增益分别从84.31%、99.04%、90.80%和32.37%提高到100%、100%、100%和100%，将假阳率从66.67%降低到0。主要原因是本文主要基于数据自身的特征（如目标模型对数据的输出概率以及数据样本间的距离）来对目标数据集中的样本进行筛选，并利用满足假阳率和成员增益条件的目标样本来实施攻击。依据的原理是模型对训练数据和测试数据的表现不同，模型更容易记住训练数据，进而在训练数据上表现更好，在测试数据上表现较差。已有工作[17,19]表明，目标数据集

中存在和成员样本相类似的非成员样本，其大大增加了攻击的假阳率。同时，MIA 通常仅能获取目标模型输出结果，而无法访问原始数据标签。MMD 距离的无监督特性使其特别适用于此类攻击。此外，尽管其他距离度量指标（如 Kullback-Leibler 散度），也可以用来测量分布差异，但由于存在计算不稳定或对噪声敏感等问题，使其在高维空间和复杂分布场景中表现不佳。因此，本文利用 MMD 距离可准确评估不同样本间的相似性，而样本间的 MMD 距离越大，降低了样本区分难度，从而增强攻击效果。

本文提出的成员推理攻击有效性检测算法虽可检测已有 MIA 的有效性。但由于不同攻击者对假阳率的可接受能力不同，因此该算法设置的假阳率阈值 α 和成员增益阈值 ϖ 很难统一，对于不同的攻击者需要进行多次测试。此外，基于数据特征的成员推理增强攻击方法主要利用数据特征筛选目标样本，这些数据特征包括目标模型对数据的输出概率以及数据样本间的距离，其并没有涉及对数据处理后提取的特征。因此，该方法仅能提升基于模型输出概率的攻击性能，而对于基于其他原理的攻击的增强效果仍不清楚。同时，当多个攻击者（如 l 个）利用该方法来增强攻击性能时，由于不同场景下不同攻击者对假阳率要求不同。因此，当其达到假阳率要求时需测试的次数分别为 $\{\mathrm{Num}_1, \mathrm{Num}_2, \cdots, \mathrm{Num}_l\}$，那么完成一次增强攻击所需的测试次数为这 l 个次数之和。

5 结束语

针对已有工作未充分检测现有成员推理攻击方法的有效性和实用性，导致错误推测误导攻击者的判断，本文首先考虑数据特征设计了一种成员推理攻击有效性检测算法，并检测出现有攻击方法的假阳率高达 100%。然后，结合攻击的基本原理，提出了一种基于数据特征的成员推理增强攻击方法，在降低攻击假阳率的同时提高攻击效率。本文方法在多个基准数据集上，可将 15 个现有 MIA 方法的假阳率降低到 0，同时将攻击精确率和成员增益都提高到 100%，实现了攻击效果的增强。

参考文献：

[1] 郭虎升, 孙妮, 王嘉豪, 等. 基于自适应深度集成网络的概念漂移收敛方法[J]. 计算机研究与发展, 2024, 61(1): 172-183.
GUO H S, SUN N, WANG J H, et al. Concept drift convergence method based on adaptive deep ensemble networks[J]. Journal of Computer Research and Development, 2024, 61(1): 172-183.

[2] 廖鹏, 方滨兴, 刘潮歌, 等. NFT 仿冒欺诈的测量与检测技术[J]. 计算机学报, 2024, 47(5): 1065-1081.
LIAO P, FANG B X, LIU C G, et al. Measurement and detection techniques for NFT counterfeiting fraud[J]. Chinese Journal of Computers, 2024, 47(5): 1065-1081.

[3] SHOKRI R, STRONATI M, SONG C Z, et al. Membership inference attacks against machine learning models[C]//Proceedings of the 2017 IEEE Symposium on Security and Privacy (SP). Piscataway: IEEE Press, 2017: 3-18.

[4] LIU H, WU Y H, YU Z Y, et al. Please tell me more: privacy impact of explainability through the lens of membership inference attack[C]//Proceedings of the 2024 IEEE Symposium on Security and Privacy (SP). Piscataway: IEEE Press, 2024: 4791-4809.

[5] SALEM A, ZHANG Y, HUMBERT M, et al. ML-leaks: model and data independent membership inference attacks and defenses on machine learning models[C]//Proceedings 2019 Network and Distributed System Security Symposium. Internet Society, 2019: 1-15.

[6] YEOM S, GIACOMELLI I, FREDRIKSON M, et al. Privacy risk in machine learning: analyzing the connection to overfitting[C]//Proceedings of the 2018 IEEE 31st Computer Security Foundations Symposium (CSF). Piscataway: IEEE Press, 2018: 268-282.

[7] HUI B, YANG Y C, YUAN H L, et al. Practical blind membership inference attack via differential comparisons[C]//Proceedings 2021 Network and Distributed System Security Symposium. Internet Society, 2021: 1-17.

[8] NASR M, SHOKRI R, HOUMANSADR A. Machine learning with membership privacy using adversarial regularization[C]//Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security. New York: ACM Press, 2018: 634-646.

[9] 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. New York: ACM Press, 2016: 308-318.

[10] JIA J Y, SALEM A, BACKES M, et al. MemGuard: defending against black-box membership inference attacks via adversarial examples[C]//Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security. New York: ACM Press, 2019: 259-274.

[11] SHEJWALKAR V, HOUMANSADR A. Membership privacy for machine learning models through knowledge transfer[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2021, 35(11): 9549-9557.

[12] MAZZONE F, HEUVEL L D, HUBER M, et al. Repeated knowledge distillation with confidence masking to mitigate membership inference attacks[C]//Proceedings of the 15th ACM Workshop on Artificial Intelligence and Security. New York: ACM Press, 2022: 13-24.

[13] JARIN I, ESHETE B. MIAShield: defending membership inference attacks via preemptive exclusion of members[J]. Proceedings on Privacy Enhancing Technologies, 2023, 2023(1): 400-416.

[14] WANG X D, WU L F, GUAN Z T. GradDiff: gradient-based membership inference attacks against federated distillation with differential comparison[J]. Information Sciences, 2024, 658: 120068.

[15] SHACHOR S, RAZINKOV N, GOLDSTEEN A, et al. Improving membership inference attacks against classification models[C]//International KES Conference on Intelligent Decision Technologies. Singapore: Springer Nature Singapore, 2025: 169-179.
[16] HE X L, XU Y, ZHANG S C, et al. Enhance membership inference attacks in federated learning[J]. Computers & Security, 2024, 136: 103535.
[17] CARLINI N, CHIEN S, NASR M, et al. Membership inference attacks from first principles[C]//Proceedings of the 2022 IEEE Symposium on Security and Privacy (SP). Piscataway: IEEE Press, 2022: 1897-1914.
[18] SONG L W, MITTAL P. Systematic evaluation of privacy risks of machine learning models[C]//Proceedings of the USENIX Security Symposium. Berkeley: USENIX Association, 2025: 1-19.
[19] WATSON L, GUO C, CORMODE G, et al. On the importance of difficulty calibration in membership inference attacks[J]. arXiv Preprint, arXiv: 2111.08440, 2021.
[20] YE J Y, MADDI A, MURAKONDA S K, et al. Enhanced membership inference attacks against machine learning models[C]//Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security. New York: ACM Press, 2022: 3093-3106.
[21] GRETTON A, BORGWARDT K M, RASCH M J, et al. A kernel two-sample test[J]. The Journal of Machine Learning Research, 2012, 13(1): 723-773.

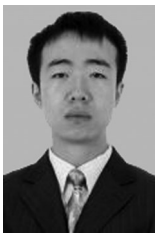
[作者简介]

牛俊（1992– ），女，陕西西安人，博士，西安电子科技大学在站博士后，主要研究方向为人工智能安全。

沈括（1999– ），男，宁夏中卫人，西安电子科技大学硕士生，主要研究方向为人工智能安全。

王月琮（1999– ），女，甘肃兰州人，西安电子科技大学博士生，主要研究方向为通信及人工智能安全。

韩雪雪（1994– ），女，河北石家庄人，西安电子科技大学博士生，主要研究方向为网络安全。



张嘉伟（1985– ），男，山西太原人，博士，西安电子科技大学讲师、硕士生导师，主要研究方向为数据安全、无线网络安全、AI 安全、区块链和云安全。

李兴华（1978– ），男，河南南阳人，博士，西安电子科技大学教授、博士生导师，主要研究方向为无线网络安全与隐私保护、智能无人系统安全。

张玉清（1966– ），男，陕西宝鸡人，博士，中国科学院大学教授、博士生导师，主要研究方向为网络与系统安全。