Understanding Membership Inferences on Well-Generalized Learning Models

一、Abstract

以往的研究表明，当模型被过度拟合到训练数据上或当对手控制训练算法时，MIA是可行的。但是，当没有过拟合且敌手没法控制算法时，MIA的威胁是未知的。

本文发现过拟合是MIA成功的充分条件而非必要条件，提出了一种新的generalized MIA（GMIA），GMIA可以找到脆弱样本并评估它们被过拟合指标忽略的影响，而脆弱样本可以通过查询其他相关样本来间接受到攻击，现有泛化技术无法很好地避免这种问题。

二、Introduction

本文重新审视了MIA的威胁，试图回答了以下问题：（虽然没有完全解决以上问题，但离答案更进了一步。）

**（1）过度拟合是成员泄露的根本原因吗?**

过拟合是充分非必要条件（实验中得到了证明）。

另外，还发现敌手甚至不需要查询模型，而使用相关的其他样本及它们的分类结果来确定目标样本是否为成员。

**（2）泛化是否是解决方法？**

现有正则化无法阻止GMIA，即便是L2范式。

**（3）那么成员泄露根本原因是？**

成员泄露是由于训练集里的一些样本对模型有特别的影响，进而影响了模型输出。泛化方法可以抑制训练样本引入的噪声，但不能完全消除它对输出的影响。差分隐私引入小噪声降低了样本的影响，但是却损失了模型精度。因此解决方式仍待研究。

本文提出的GMIA思想如下：

GMIA在训练集中找到脆弱样本，并检测了它们之前被过拟合指标忽略的影响。

对泛化良好的模型，成员与非成员表现差异不大，无法再使用shadow model实施攻击。因此GMIA检测与分析脆弱样本（outliers），推断是否为成员。

三、Background

p-values： p值是报告研究结果时发布的一种汇总统计数据，可用于成功推断成员。

泛化误差：低泛化误差意味着良好的泛化能力，会在训练与测试数据上表现相近。实践中， 由于接触不到population,可以用它的hold out集测量泛化误差。

敌手能力设置：黑盒访问权（但问询次数有限制）或指导目标模型的结构与训练算法；可以访问独立于population的一组记录，这组记录可能与训练集是重叠或非重叠的（敌手对此不知道）。

**四、Un****derstanding Membership Inference Attacks**

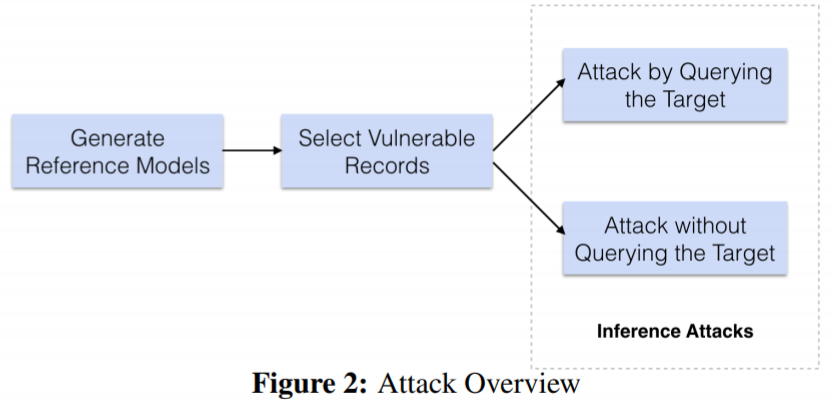
对大量模型来说，当记录r在或不在训练集中时，记录q的输出会不同，那么，r对q就有独特影响。r即为要找的脆弱样本。

注意，重要的不是记录的影响强度，而是它对训练集中的其他记录有着独特影响。如果训练集中有记录对模型的影响与r相似，r就没有独特影响，那么攻击就失败了。

实验表明，用这种间接推理的方式有时比直接推理效果更好，因为敌手可以从多个查询中积累证据，查询越多，越有可能搜集到目标记录的独特影响。

**五、Generalized Membership Inference Attack**

GMIA框架如下：



如图，GMIA包括了三个步骤：参考模型的生成、选择脆弱记录、推理攻击。推理攻击又分为直接推理与间接推理。

每一步概述如下：

**·Generate Reference Models**

建立一个参考模型来模仿目标模型的预测行为。因为可用的参考记录有限，用bootstrap采样构造训练集，并用其训练多个参考模型。构建好后，参考模型在GMIA框架中的每个步骤都会得到使用。

**·Selecting Vulnerable Target Records**

对泛化良好的模型，不是所有训练数据都是脆弱记录。本文想到了一种找出脆弱记录的方式，即估计记录在参考数据代表的样本空间中的邻居数量，邻居较少的记录可能对模型有独特影响，因此更易受到MIA攻击。根据每个记录在参考模型上的中间输出，为该记录构造一个新特征向量表示邻域，向量同时也反映了记录对模型的影响。

**·****Direct Inference by Querying the Target Record**

直接推理情况下，使用对目标记录的问询实施攻击。当目标记录不是成员时，用假设检验判断目标模型的预测是否偏离了参考模型的预测。假设检验中的p值是攻击的置信度。

**·****Indirect Inference without Querying the Target Record**

训练数据不仅会影响模型对自己的输出，还会影响其他看似不相关的记录（这里称之为enhancing records）。用新的方式迭代找到enhancing records，使用enhancing records可间接推断出目标记录的存在而无需直接查询目标模型。

每步骤更多细节如下：

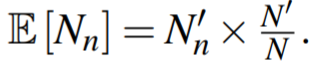
**·Generate Reference Models**

使用bootstrap sampling产生参考数据集可以使各参考数据集之间的重叠尽量小。参考数据集与目标训练集的采样空间是一样的，但是不包含目标记录。然后每个参考数据集训练一个参考模型，每个参考模型的训练算法是一样的。

**·Selecting Vulnerable Target Records**

测量目标记录的潜在影响，从而选择影响最大的脆弱记录。注意，虽然对模型产生独特影响的常为训练集中的离群记录outlier records，但反之不成立，并非outliers一定会产生独特影响，原因在于训练算法会给一些特征更高的权重。因此这里不用输入特征，而提取相关的high level特征来检测脆弱记录。

具体的讲，攻击神经网络时，将参考模型的Softmax函数之前的最后一层输出连接起来，构造新的特征向量，用新的特征向量测量每条记录的独特影响。当两个记录的特征向量的余弦距离小于阈值时，我们说它俩是邻居。本文MIA的假定目标是训练集中有更少甚至没有邻居的脆弱记录。

训练集有N条记录，参考训练集有条记录（两个训练集从同一空间采样），目标记录为r，计算r在参考数据集中的邻居数为，在r在训练集中邻居数可以用下式估计：。当r时（β是阈值），r被认为是要找的脆弱记录。

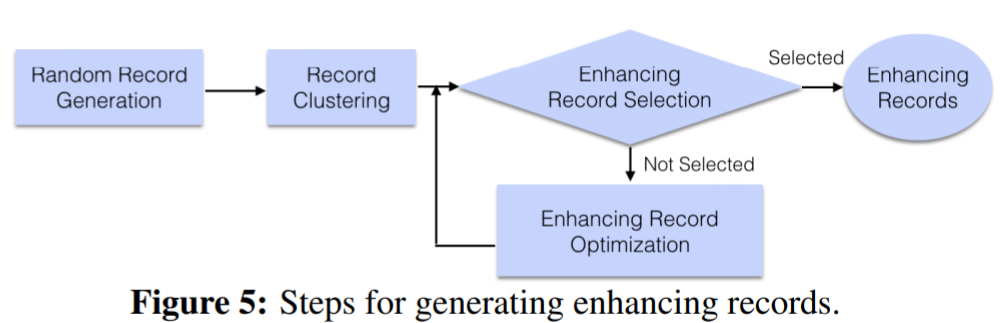
**·Direct Inference by Querying the Target Record**

对每个目标记录，计算目标模型给的输出与参考模型给的输出之间的偏差进行攻击。用log损失函数量化输出间差异，当模型对记录的正确标签给出了高概率时，log损失函数会很小。p值与log损失之间有函数关系（文中用的是empirical cumulative distribution function），当p值小于阈值（如0.01），说明r在目标模型训练集中。

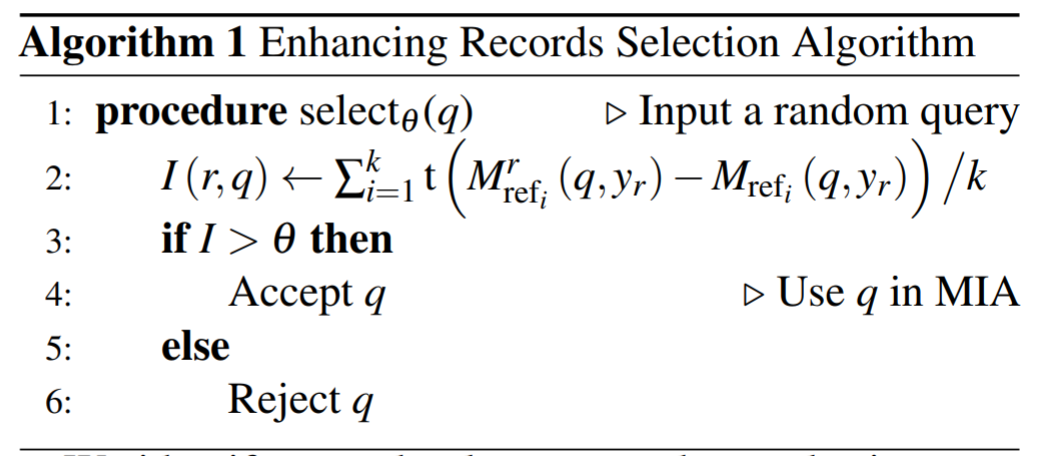
**·Indirect Inference without Querying the Target Record**

MIA可以查询看似与目标记录不相关的记录实现攻击，这样攻击很难被检测到。

间接攻击的关键在于找到增强记录，增强记录是那些输出受到目标记录影响的记录。找到增强记录的步骤如下图：



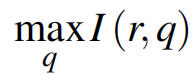
1. 生成随机记录。之后的增强记录会在这些记录中。当特征空间较小时，采均匀采样得到随机记录；当特征空间较大时，在预先选择的脆弱目标记录中加入高斯噪声生成随机记录。
2. 选择增强记录。用一组positive参考模型近似目标模型行为，这些positive参考模型训练集为参考记录加目标记录r。（为减小开销，将目标记录添加到从原始参考数据集采样的batches里，对batches和r训练更新参考模型。）



第二行计算目标记录r对q的影响，当影响大于阈值时，p被认为是r的增强记

录。

3. 增强样本优化。当目标模型样本空间很大时，在随机记录中找到增强记录的机会

很小。这时使用优化方式找增强记录，对下式进行优化，其中I是上面

第二行的影响。

4. 记录聚合（可选）。为提高查询效率，提出一种从大量随机记录中识别出相关性最

小的增强记录的算法。简单地说就是先根据模型预测估计记录间的相关性，将选

择最小相关记录子集问题视为图论中的k-lightest 子图问题，用层聚类得到近

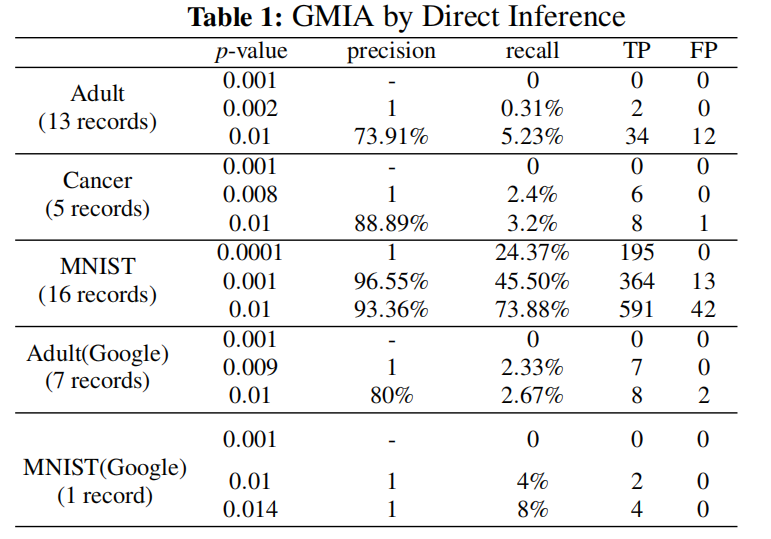
似解。这一步可选，放在2,3之前。

5. 有多个查询的间接推断。识别多个增强记录后，查询每个记录，重复之前详述的直

接推断攻击。

**六、Experiment**

1、GMIA直接推断攻击准确率如下：

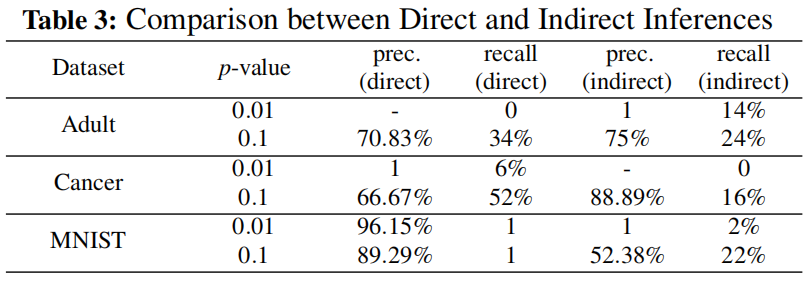


使用的数据集都是泛化良好的（训练集和测试集精度差异小于0.01）。

前三行都是分别使用这三种数据集在本地训练的模型，GMIA攻击精度较高，说明了即便是没有过拟合的模型存在很高的隐私泄露风险。

最后一行GMIA效果不那么好的原因可能在于无法获得准确的模型结构。

2、GMIA间接攻击效果、与直接攻击效果对比如下：



尽管间接攻击在MNIST数据集上不如前两个数据集好，但也证明了在不查询目标

记录的情况下，也可以成功推断出成员与非成员。

**七、Mitigation**

1、应用正则化与差分隐私（加扰动）时，同时进行适当的训练集选择、检测与删除脆弱样本。

2、自动识别outliers并删除对模型效用贡献不大的值。