AUTOLYCUS：利用可解释人工智能（XAI）对可解释模型发起模型提取攻击*†

Abdullah Caglar Oksuz 凯斯西储大学 美国俄亥俄州克利夫兰 abdullahcaglar.oksuz@case.edu

Anisa Halimi IBM Research - Dublin 爱尔兰都柏林 anisa.halimi@ibm.com

Erman Ayday 凯斯西储大学 美国俄亥俄州克利夫兰 erman.ayday@case.edu

arXiv:2302.02162v3 [cs.LG] 8 Jul 2024

摘要

可解释人工智能（XAI）旨在揭示人工智能模型的决策过程。然而，用于生成此类解释的数据可能带来安全与隐私风险。现有文献指出，机器学习模型会受到多种攻击，包括成员推断、模型反演和模型提取攻击。根据具体设置及相关参与方，这些攻击的目标可能是模型，也可能是训练数据。

可解释人工智能工具会增加模型提取攻击的风险。当模型所有者倾向于采用黑盒访问，从而对模型参数和架构保密时，这一点尤其值得关注。为利用这一风险，我们提出了 AUTOLYCUS，这是一种新颖的、基于重新训练（学习）的模型提取攻击框架，旨在黑盒设置下攻击可解释模型。作为可解释人工智能工具，我们利用局部可解释模型无关解释（LIME）和 Shapley 值（SHAP）来推断决策边界，并创建能够复现目标模型功能的替代模型。选择 LIME 和 SHAP，主要是因为它们能提供逼真且信息丰富的解释，同时又具有广泛的应用、简单易用等特点。

我们在六个机器学习数据集上评估了 AUTOLYCUS，衡量替代模型相对于目标模型的准确率和相似度。结果表明，AUTOLYCUS 效果显著，所需查询次数远少于最先进的攻击，同时准确率和相似度与之相当。我们在多种可解释机器学习模型上验证了其性能和可迁移性，包括决策树、逻辑回归、朴素贝叶斯和 K近邻。此外，我们还展示了 AUTOLYCUS 对所提出的对策具有韧性。

关键词

模型提取攻击、可解释人工智能、可解释人工智能、机器学习中的隐私攻击、自适应再训练

1 引言

机器学习（ML）已成为许多企业的重要工具，使它们能够分析数据，并以更高的准确率和效率进行预测。为了让更多企业能够使用机器学习，云服务提供商推出了经济实惠的机器学习即服务（MLaaS）平台，为客户提供预训练模型或部署自有模型的工具。

随着机器学习即服务平台的普及，对促进可解释人工智能（XAI）的工具的需求也相应增加 [4]。这些工具对于提供用户透明性，并帮助他们全面了解机器学习模型如何做出决策至关重要 [7]。对于本质上以黑箱方式运行的模型（例如神经网络），或是在黑盒设置下运行的可解释模型（这正是本文的重点研究对象，即用户无法或只能有限地访问模型内部运作方式），其缺乏可解释性和透明性可能会引发不信任，并阻碍其推广应用。

局部可解释模型无关解释（LIME） [41]和 SHapley 加性解释（SHAP）是两种广受欢迎的可解释人工智能技术，主要由 Azure 机器学习（负责任人工智能） [2], 、Watson 机器学习（OpenScale） [11], 、Amazon SageMaker（Clarify） [47], 和 Google Cloud（Vertex 可解释人工智能） [10]等机器学习即服务平台采用。LIME 通过围绕感兴趣样本创建一个新模型，为任意机器学习模型预测生成局部解释。而 SHAP 则利用合作博弈论 [37] 中的沙普利值 [33] 以及模型的局部近似，计算特征重要性值，从而得出整个数据集的特征重要性。

机器学习即服务的商业化以及对其所提供模型的日益依赖，为企业带来了新的风险。其中一种风险是推理攻击的威胁，包括成员推断 [51], 、模型反演 [16], 和模型提取 [55]。具体而言，Tramer 等人 [55] 提出的针对机器学习即服务平台的模型提取（或窃取）攻击，可能导致专有模型和宝贵知识产权被盗，并造成经济收益损失。在此类攻击中，对手会尝试通过向目标模型发送一系列查询，并利用查询结果训练代理模型，从而提取全部或部分目标机器学习模型。此外，攻击者还可以利用可解释人工智能工具提供的信息，更轻易地实施模型提取攻击。例如，Milli 等人 [34] 提出了首个利用解释、借助基于梯度的解释求解神经网络图像分类器超平面的模型提取攻击。Chandrasekaran 等人 [8]证明，模型提取攻击可以表述为主动学习。除了隐私问题之外，Rudin [44] 提出，对于高风险决策，应更多地采用可解释模型，而不是解释黑箱（不可解释）模型。这有助于确保问责制，并促进人们对人工智能系统的理解与信任。

尽管利用人工智能解释实施对抗性机器学习攻击还是一个较新的研究领域，但已有多项研究 [24, 34, 35, 56, 59]在可解释人工智能的指导下，针对神经网络和图像分类任务开展模型提取攻击。这些研究主要依赖图像分类器专用解释工具（例如类别激活图（CAM） [46, 63]），与模型无关的解释相比，这类工具能提供更丰富的基于梯度的信息。

*发表于《隐私增强技术会议论文集》（PoPETs），第 2024 卷，第 4 期，2024 年。

†源代码请参阅 https://github.com/acoksuz/AUTOLYCUS。

与模型无关的解释。尽管这些研究颇有启发性，但它们只关注神经网络和图像分类器，忽视了探索其他模型架构和多样化数据设置可能带来的见解与进展。仅依赖适用于神经网络的解释方法，可能与某些模型类型不兼容，或无意间泄露过多关于模型的信息；对于复杂度较低的可解释模型而言，这实际上会使攻击转变为半白盒场景。相比之下，我们的目标是以尽可能小的干扰，对多个可解释模型实施模型无关攻击。此外，针对可解释模型的研究仍然有限，而这类模型因其简单性，至今仍广泛应用于常规统计建模和数据分析。因此，探索人工智能解释对这类模型构成的漏洞，尤其是在黑盒设置下，仍是一片未被探索的领域。

在这项工作中，我们提出了一种新颖的、模型无关的、基于再训练（学习）和扰动的模型提取攻击框架 AUTOLYCUS。该框架利用可解释人工智能（XAI），在黑盒访问设置下针对可解释模型发起攻击。作为一种先验引导的对抗样本构造方法，AUTOLYCUS 利用模型对局部决策边界的解释和特征重要性。SHAP 提供特征重要性，而 LIME 同时提供特征重要性和决策边界。攻击者可以利用这些信息生成成员查询。

AUTOLYCUS 受到 Tramer 等人提出的线搜索再训练和自适应再训练方法的启发；这两种方法被指定为仅标签场景下的主要策略 [55]。线搜索再训练会根据先前的查询，向模型决策边界附近发送自适应成员查询，这与其他仅标签攻击采用的均匀查询不同 [9, 38]。AUTOLYCUS 在解释的利用方式上不同于上述攻击。AUTOLYCUS 不会直接寻找决策边界，而是在解释附近进行线搜索，以生成能够在再训练过程中准确反映决策边界的样本。虽然 AUTOLYCUS 可以利用额外数据集或关联数据等外部资源作为对手的先验信息，但我们将访问权限限制为黑盒访问，并约束解释所披露的信息。方程求解和路径搜索等方法 [55] 会受到辅助白盒模型信息和真实性方面的限制；相比类似工作 [34, 38, 55]，我们的方法对对手的假设要求更宽松，同时兼顾实用性和真实性。

我们通过构建的代理模型的准确率及其与目标模型的相似度，展示了 AUTOLYCUS 在不同机器学习数据集上的有效性。我们表明，AUTOLYCUS 对决策树、逻辑回归、朴素贝叶斯和 K 近邻等可解释模型，均能实现较高的准确率和相似度。此外，我们还展示了 AUTOLYCUS 优于最先进方法，并且即使存在对策也依然有效。

贡献 AUTOLYCUS 的主要贡献如下。

- AUTOLYCUS 所需的启动条件极少（例如，只需少量超参数知识和辅助数据），因此是一种易于部署且高效的攻击。
- 与精确重构攻击（如方程求解、 [55], 路径搜索 [55], 和 Lowd-Meek [31]）相比，• AUTOLYCUS 能够以更少的查询次数，生成保真度和准确率都很高的目标模型重构。AUTOLYCUS 是一种基于再训练的攻击。即使受到严格的查询限制，它也能创建部分提取的替代模型，其准确率和相似度都相当高。
- 据我们所知，AUTOLYCUS 是首个利用人工智能解释、通过黑盒访问来攻击可解释模型的模型提取攻击。
- 据我们所知，AUTOLYCUS 是首个利用模型无关的人工智能解释工具的模型提取攻击。

本文其余部分安排如下。下一节将回顾相关工作。第3节介绍可解释人工智能的背景。第4节描述所考虑的系统与威胁模型。第5节详细介绍 AUTOLYCUS 的系统模型。第6节给出评估结果。第7节展示存在潜在对策时 AUTOLYCUS 的性能。第8节讨论潜在扩展和局限性。最后，第9节总结全文。

2 相关工作

我们回顾两类主要的相关研究：（i）机器学习中的隐私攻击，以及（ii）由模型解释带来的隐私风险。**机器学习中的隐私攻击**。机器学习隐私方面的研究发现了多种攻击，例如，成员推断 [9, 30, 45,51]，用于确定某个用户是否属于训练数据集；属性推断 [18, 53]，根据观察到的用户属性推断其其他隐私属性；模型反演 [6, 16, 61]，用于从训练数据集中重构数据；以及模型提取 [55]，用于推断模型的参数/超参数和/或架构。在这项工作中，我们关注模型提取（或窃取）攻击，即攻击者试图从训练好的模型中提取敏感信息或知识产权。攻击者通常可以访问模型的输入输出行为，并利用这些信息构建模型副本（功能提取）[36, 49]，或从中提取其他有价值的信息（属性提取）[13, 22, 55, 57]。Tramer 等人 [55] 提出了首个同时考虑白盒和黑盒访问机器学习模型的模型窃取攻击。Wang 等人 [57] 提出了一种针对机器学习模型超参数的模型提取攻击。Papernot 等人 [38] 提出了一种针对黑盒深度神经网络分类器的模型提取攻击。这类攻击对多种机器学习模型均有效，包括决策树、深度神经网络和支持向量机 [21, 34, 55, 57]。

模型解释带来的隐私风险。虽然可解释人工智能技术有助于理解机器学习模型提供的预测，但也可能被用于加强隐私攻击。Shokri 等人 [50] 研究了基于特征的模型解释的成员推理攻击所带来的隐私风险，并展示了各种解释方法如何泄露有关训练数据的信息。Zhao 等人 [62] 分析了解释方法如何增强模型反演攻击。他们的工作聚焦于 MNIST [29] 和 CIFAR-10 [28], 等图像数据集，并表明模型反演攻击的准确率会显著提高

在使用解释时。Kariyappa 等人 [24], Milli 等人, [34], Miura 等人 [35], Truong 等人 [56], 和 Yan 等人 [59] 提出了多种模型提取攻击，通过在黑盒图像分类器上利用基于梯度的人工智能解释（例如显著图 [52], 、类别激活图 [63]）来实施攻击。Aïvodji 等人 [1] 利用反事实解释实施模型提取攻击。本文不同于 Shokri 和 Zhao 的工作，聚焦于利用解释实施的模型提取攻击。

3 可解释人工智能背景

可解释人工智能（XAI）是一种用于开发人工智能系统的方法，旨在使其具备透明性、可解释性，并能够为其决策过程提供清晰的解释 [12, 19]。XAI 的目标是确保人工智能系统值得信赖、公平且负责任，并能够以可靠且易于理解的方式为人类决策提供参考。传统人工智能算法（如深度学习神经网络）能够基于大量数据高效地做出决策，但由于其内部运作机制难以理解，通常被视为“黑箱”。这可能引发人们对决策过程中偏见、歧视或错误的担忧 [44]。XAI 旨在通过将透明性和可解释性技术融入人工智能系统的设计来解决这些问题，包括那些已被认为具有可解释性的系统。这可能涉及使用更简单的模型、提供可视化或自然语言解释，或检测异常行为。

在相关文献中，研究人员开发了各种可解释人工智能技术，以提高机器学习模型的可解释性和透明性，尤其是计算机视觉领域中的神经网络模型。这些技术包括逐层相关性传播 (LRP) [3],深度泰勒分解 (DTD) [27],模式差异分析 (PDA) [64], 概念激活向量测试 (TCAV) [26], 以及可解释图神经网络 (XGNN) [60]。对于决策树等可解释模型，则可以使用 Anchors [42], SHAP [33], 和 LIME [41] 等可解释人工智能技术来提供解释。

有关主要机器学习即服务平台所采用技术的更多信息，以及本文为何主要关注 LIME 和 SHAP，请参阅附录。

局部可解释模型无关解释 (LIME) LIME[41]是模型所有者广泛使用的可解释人工智能工具，用于理解机器学习模型预测背后的依据。它通过为感兴趣的特定实例或样本生成模型周围的局部近似来实现这一点。例如，要为模型 $\hat{M}$ 上的样本 X 提供解释，LIME 会创建一个数据集，其中包含不同的 X，这些都是 X 的扰动版本。扰动的执行方式因数据类型而异，例如对图像数据进行像素遮蔽，或对文本数据进行词语替换。随后， M 会对扰动数据集中的每个样本进行分类。接着，LIME 会围绕 X 拟合一个局部线性模型，使用 $\hat{X}$s 及其由 M 给出的分类。由此得到的线性模型具有自身的局部决策边界，可用于解释 M 的线性和非线性边界，从而允许在其邻域内生成并利用样本。局部线性模型中的特征权重表示每个特征在 M 的预测中的相对重要性，并通过热力图或特征重要性分数等方式呈现。需要注意的是，由于非线性的存在，LIME 生成的局部线性模型的决策边界可能与 M的确切决策边界并不完全吻合。但是，对于可解释模型而言，与为神经网络等黑盒模型提供的解释相比，这些局部解释更加准确且一致。

沙普利值 (SHAP) SHAP 利用样本的沙普利值 [48] 来确定每个特征的边际贡献。随后，这些值会作为针对给定预测的该样本的解释呈现出来。SHAP 所披露的信息可能因预测而有很大差异。例如，它可能仅披露每个类别对应的沙普利值——考虑到黑盒访问，这是最现实的情形；也可能提供一张图，展示训练数据集中的每个样本以及该特定样本。不用说，后一种方式会泄露大量有关模型的信息。在 AUTOLYCUS 中，我们假设采用前一种方式。要将 SHAP 集成到 AUTOLYCUS 中，我们需要将沙普利值转换为决策边界（如 LIME 所做的那样），或者仅根据这些值计算特征重要性。鉴于 SHAP 解释具有无向且标量的结构，无法推导出决策边界。因此，我们采用 SHAP 的特征重要性。

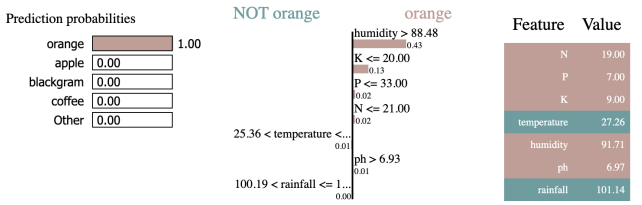


图1：Crop 数据集中的 LIME 解释示例

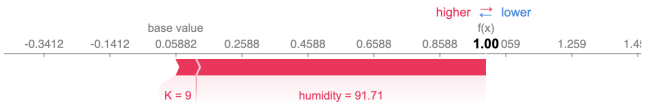


图2：Crop 数据集中的 SHAP 解释示例

我们展示了将 LIME 和 SHAP 应用于在 Crop 数据集 [20] 上训练的决策树模型的示例。图1展示了 LIME 对一个成员查询的解释，该查询被目标模型分类为橙色。左侧显示了各类别的预测概率。图的中间部分列出了促成模型决策的局部因素。这些因素按照解释力从强到弱排序。每个局部因素由促成该因素的特征及其对应的决策边界区间构成。右侧列出了特征及其取值，并以颜色标示这些特征是增强还是削弱模型的预测。图2展示了对同一样本的 SHAP 解释及特征重要性。这些示例旨在说明 LIME 和 SHAP 的功能，以及它们为可解释模型提供的信息类型。关于 Crop 数据集和实验设置的更多细节，请参见第6.1节及后续章节。

4 系统与威胁模型

考虑一个在数据集 D_M 上训练的目标模型 M，该模型通过 API 部署供用户使用（例如，在机器学习即服务（MLaaS）平台上）。用户可以向目标模型 M 发送查询，并收到作为输出的预测标签，以及由可解释人工智能（XAI）工具生成的相应解释（如第 3 节所述）。例如，用户向目标模型 M 发送针对给定样本 X_i 的查询，并收到包含预测类别 y_i 及其对应解释 E_i 的输出，这些解释由 XAI 提供（如图3所示）。当对手试图利用预测标签的输出以及 XAI 工具返回的解释，学习一个能够紧密近似目标模型 M 的代理模型 S 时，就会发生模型提取攻击。攻击者的目标是创建一个能够内在反映决策边界、进而反映目标模型 M 功能的代理数据集。

我们假设攻击者拥有对目标模型 M 的黑盒访问权限，包括了解机器学习模型的类型（而非架构），例如决策树、逻辑回归。即使模型类型未知，具备机器学习可迁移性专业知识的个人也能在重新训练期间以合理的准确率推断出模型类型 [39]。此外，攻击者获知 y_i 可能取值，即类别名称，因此也知道类别总数 t。攻击者可以访问一个辅助数据集 D_A，其中包含一个或多个样本，用于初始化遍历算法。这是为了避免发送盲目查询，从而造成查询预算的额外开销。 D_A 的组成可以从简单的样本集合到全面的数据集不等。请注意，即使没有样本，仍可以使用默认样本值进行随机查询来启动该过程。这可用于模拟攻击者缺乏辅助信息的情形。然而，样本有限会限制探索能力；与可用更全面的 D_A 的情形相比，要高效地实施模型提取攻击，就需要更高的查询预算。最后，攻击者知道用于训练模型 M 的数据集中特征的基本属性，例如其类型、取值域以及下界和上界（即 $0 \leq \text{pH} \leq 14$）。

请参考表 1，其中列出了本文中经常出现的符号和记号。

表 1：常用符号和记号。

M	目标模型
S	代理模型
D_M	目标模型的训练数据集
D_S	代理模型的训练数据集
D_A	攻击者的辅助数据集
D_E	探索数据集
Q	查询预算（查询数量）
X_i	样本 i
$\hat{X}_i$	修改后的样本 i
x_i^j	样本 i的 j^{th} 特征
$\hat{x}_i^j$	样本 i的修改后特征 j
y_i	样本 i的预测类别
E_i	样本 i的解释
Exp_M	模型 M的解释器函数
k	允许探索的排名靠前的特征数量
m	特征总数
n	每个类别的样本数
t	类别总数
l, u	每个类别生成的样本数的下界和上界
db_i^j	样本 i的 j^{th} 特征的决策边界
δ	变更系数
TRAV-A	遍历算法

5 提出的方法

在 AUTOLYCUS 中，攻击者首先需要创建一个代理数据集 D_S，以训练出能够紧密近似目标模型 M 的代理模型 S。我们假设攻击者可以访问辅助数据集 $D_A = X_1$ 中的 n 个每类样本，$X_2, \ldots, X_{t*n}$ 这些样本可能包含也可能不包含来自原始数据集 D_M 的样本。这里，$X_i = \{x_i^1, x_i^2, \ldots, x_i^m\}$，其中 x_i^j 表示样本 X_i 中特征 j 的取值，而 m 是特征总数。我们还假设存在查询预算 Q，用于限制攻击者向目标模型 M 发送的查询总数。所提出的模型提取攻击如图 3 所示。在步骤 1 中，攻击者使用一个局部样本（来自辅助数据集）发送查询，以生成额外样本。在步骤 2 中，确认查询的有效性。在步骤 3 中，机器学习模型 M 预测被查询样本的类别，LIME 或 SHAP 则计算其解释。在步骤 4 中，机器学习即服务平台返回模型对类别的预测及其解释。在最后一步（步骤 5）中，攻击者解析并利用模型解释来生成修改后的样本，这些样本与其前序样本之间仅有一个或 k 个特征不同。SHAP 会返回 M 所预测类别中每个特征的沙普利值，并将其解析为特征重要性。与发送随机查询相比，这一过程由解释提供信息并加以引导。随着发送的查询数量增加，会获取到更多带有目标模型预测标签的样本。在任何再训练攻击中，用作代理模型 S 训练数据集 D_S 的样本，旨在模仿目标模型的行为。下文将介绍 AUTOLYCUS 的具体细节。

5.1 生成候选样本

攻击者的目标是向目标模型 M 发送具有信息量的查询，而不是盲目查询。为此，根据目标模型所使用的可解释人工智能工具，攻击者探索 LIME解释返回的决策边界或 SHAP解释返回的特征重要性。假设攻击者向目标模型 M发送针对给定样本 X_i（例如，D_A中的某个样本）的查询。如前所述，目标模型 M 会发送预测类别 y_i 及相应的解释 E_i。如第3节所述，LIME 返回的解释由预测概率组成（由于处于黑盒访问场景，我们仅假设获取排名最高的类别） M 以及决策边界 X_i（由于处于黑盒访问场景，我们仅假设获取排名最高的类别） y_i 以及决策边界 db^j i 按特征重要性排序。而 SHAP 只返回特征重要性。为了生成新的、信息丰富的候选样本（在算法 1 中记为 D_C），攻击者仅考虑样本 X_i 中排名靠前的 k 个特征，并依据这些特征的重要性进行选择。

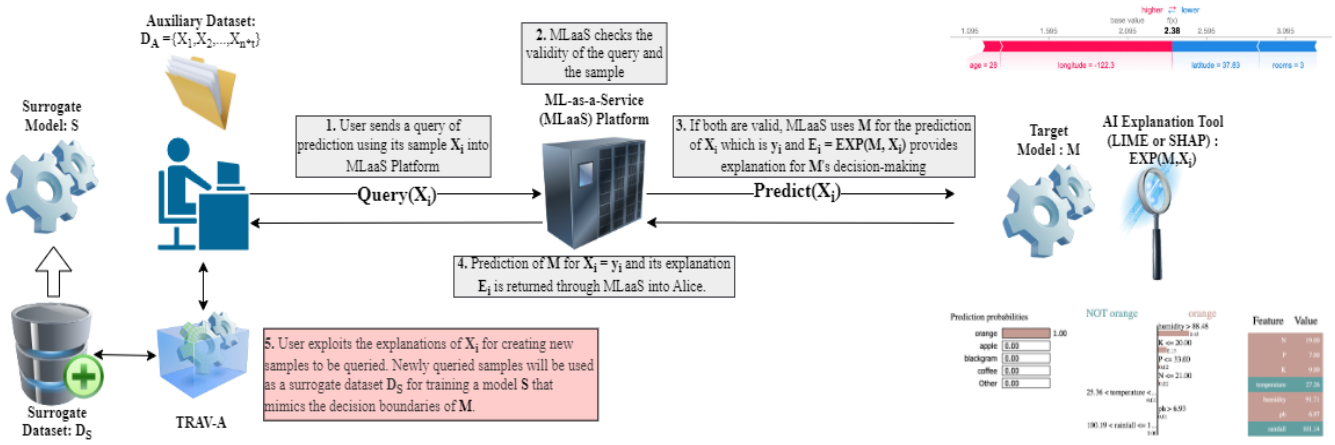


图3：AUTOLYCUS系统示意图，包含以下步骤：(1) 用户向机器学习即服务平台发送查询；(2) 机器学习即服务平台验证查询的有效性，确保不会发送空查询或不完整查询；(3) 机器学习模型M预测所查询样本 y_i的类别，解释器计算其解释 E_i；(4) 机器学习即服务平台将结果返回给用户；(5) 若用户是对抗性用户，则会利用解释，通过TRAV-A算法（如第 5.1 节所述）提取目标模型 M的决策边界。

这里有两种策略可供选择。在第一种探索过程中，特征逐一改变，从而形成广度优先树搜索；树的宽度和深度取决于所选的 k 值。在第二种探索过程中，排名靠前的 k 个特征会同时改变，因此探索到的样本集更多样。如果 D_A 的规模较小，且类别分布接近均匀分布，那么这是一种最优策略。然而，如果 D_A 的规模较大，且存在稀有类别，那么第一种方法可能会带来更好的探索效果。在第一种探索过程中，较高的 k 值会形成一棵较宽而较浅的树，这意味着需要探索更多特征以及同一样本的更多变体。这可能导致探索大量无信息的样本，并最终耗尽查询预算。相反，较低的 k 值会形成一棵较窄而较深的树，意味着检查的样本更多，同时优先考虑排名靠前的特征。这会导致对信息丰富的样本探索不足，从而可能丢失有关树结构的关键信息。选择较高还是较低的 k 值，以及将特征逐一改变还是一次性改变作为可行策略，取决于特征重要性的分布（如果已知或可推断）。

对于排名前 k 个特征中的每一个，攻击者都会分析 LIME 返回的决策边界或 SHAP 返回的特征重要性，以生成候选样本。攻击者通过将特征 j 的新值 $\hat{x}_i^j$ 修改为 LIME 中的决策边界 db_i^j，或修改为 SHAP 中下一个可用值，来计算这些新值；相应的系数记为 δ_j。形式上，$\hat{x}_i^j$ 的计算方式为：$\hat{x}_i^j = db_i^j \pm \delta_j$（在 LIME 中）或 $\hat{x}_i^j = x_i^j \pm \delta_j$（在 SHAP 中），其中 j 是攻击者试图修改的特征索引，δ_j 是修改系数。对于 LIME，类别特征的 δ 等于 1，以体现编码差异；连续特征的 δ 则为 0.01（或更低）。由于扰动由 LIME 中的决策边界决定，因此 δ 的重要性较低。另一方面，它在 SHAP 中非常重要，因为它决定了连续样本之间的探索差异。一个不错的经验法则是：如果有标准差数据，就将其设为接近标准差的值；否则设为解空间范围的四分之一。根据预期的修改方式，必要时可以手动将 δ_j 配置为更大或更小的值。由此得到的候选样本为 $\hat{X}_i = \{x_i^1, x_i^2, \ldots, \hat{x}_i^{j_1}, \ldots, x_i^m\}$ 或 $\hat{X}_i = \{x_i^1, \hat{x}_i^2, \ldots, x_i^{m-2}, \hat{x}_i^{m-1}, x_i^m\}$。

图1展示了 $X_i = \{19, 7, 9, 27.26, 91.71, 6.97, 101.14\}$，这是Crop数据集 [20] 中的一个示例样本及其对应的解释 E_i。在此例中，$k = 3$ 的排名靠前的特征是“湿度”、“K”和“P”。若聚焦于“湿度”并选择 $\delta = 0.01$，则新生成的样本将为$\hat{X}_i = \{19, 7, 9, 27.26, \mathbf{88.49}, 6.97, 101.14\}$。图2展示了SHAP对应的相同示例。在此例中，影响模型决策的唯一重要特征是“K”和“湿度”。当这两个特征都改变10（$\delta_2 = 10$ 和 $\delta_4 = 10$）时，新生成的样本将为$\hat{X}_i = \{19, 7, \mathbf{19}, \mathbf{91.71}, 27.26, \mathbf{10}, 6.97, 101.14\}$（“K”无效，因此被丢弃）

5.2 创建替代数据集

为获得平衡的数据集，攻击者旨在创建一个代理数据集 D_S，其中每个类别的样本数与目标模型的训练数据集 D_M 中相应类别的样本数相近且成比例。因此，需要生成至少达到每个类别 l 个样本、总计 $l \times t$ 个样本的最低数量，以确保 D_S 中没有任何预测类别代表不足。为高效利用查询预算 Q，攻击者将每个类别生成的样本数限制在一个上限值 u。如果不对 l 和 u 的边界值加以约束，在遍历过程中，预测频率较高的类别可能会主导 D_S，从而导致数据集失衡。这种失衡可能会损害最终替代模型 S的性能，使常见类别出现过拟合，而对较罕见类别的预测则不充分。为解决这一问题，l 和 u 值可以表示为标量

算法 1 用于生成样本的遍历算法
将用于训练代理模型

函数 TRAV-A(D_A, M, Exp_M, t, n, l, u, Q, δ**)**
  $D_E, D_S, y_S \leftarrow D_A$, {}, {}
  访问 ← {0, 0, ...0} ▷ (长度 t)
  队列 ← 0
  当 $D_E \neq \emptyset$ **且(**$Q \geq$ q **或**任意（访问））$\leq l$ **执行**
    q++
    $X_i \leftarrow D_E$.弹出()
    $y_i \leftarrow M$.预测(X_i)
    如果访问[y_i] < u **则**
      访问[y_i]++
      D_S.ADD(X_i)
      y_S.ADD(y_i)
      $E_i \leftarrow$ **解析_EXP**（X_i, M, Exp_M, t, n）
      $D_C \leftarrow$ **生成_样本**（X_i, E_i, n, δ）
      对于i **在** D_C **执行**
        如果 i $\notin D_E, D_S$，**则**
          D_E.PUSH(i)
        结束 if
      结束 for
    结束 if
  结束 while
  return D_S, y_S
结束函数

值，以便在每个类别中发现相同数量的样本；也可以表示为列表，从而允许为不同类别探索不同数量的样本。我们建议根据 D_M 中的类别频率设置 l （如果有此类信息），并通过在 l 上增加所需的误差范围来设置 u。如果类别频率未知，一个实用的经验法则是将l和 u 设置为接近 Q/t，以维持 D_S 的平衡，从而确保模型训练准确。

攻击者使用遍历算法 **TRAV-A** 来构建代理数据集 D_S。该算法通过对模型解释运行 Tramer 等人提出的线搜索重训练算法 [55] 构建而成。详细步骤见算法 1。回想一下，攻击者可以从辅助数据集 D_A中获取每个类别 n 个样本。因此，初始时，D_S 仅限于 D_A。令 D_E 表示包含待探索样本的数据集（初始时为 $D_E = D_A$）。TRAV-A 首先探索 D_E中的样本。它选取数据集 D_E 中的第一个样本 X_i，并向目标模型 M发送查询。收到预测类别 y_i后，TRAV-A 检查该类别生成的样本数是否已达到最大值。如果已达到，TRAV-A 则继续查询 D_E 中的下一个样本，并检查是否满足上述条件。否则，TRAV-A 将 X_i 添加到代理数据集 D_S，并生成新的候选样本。在第 5.1 节中，我们介绍了如何生成候选样本。将生成的、此前尚未探索过的样本（即不属于 D_E的样本）添加到 D_E中。在下一轮迭代中，TRAV-A 查询 D_E中的下一个样本。TRAV-A 采用广度优先搜索策略，遍历探索特定样本时生成的候选样本。我们采用这种方法，以防止对单个样本进行深入探索并传播。该过程会持续进行，直到满足以下任一条件：(i) D_E中没有尚未探索的样本；(ii) 查询预算 Q 已耗尽；或 (iii) 已达到每个类别需要生成的最少样本数 l。请注意，下限 l 应选取得接近每个类别的大致查询限制 Q/t，以尽可能有效地利用预算。值得注意的是，最后一个条件 (iii) 是可选的，但可补充第二个条件 (ii)。当查询预算 Q 超出，但每个类别中已探索的样本数尚不足时，便会启用这一条件。因此，如果允许超出查询预算，条件 (iii) 可替代该预算限制，并确保每个类别都有足够数量的样本得到检验。

利用生成的代理数据集 D_S，攻击者训练替代模型 S。请注意，攻击者知道目标模型的类型 M，并且可以在不同的超参数设置下创建多个替代模型。随后，攻击者无需消耗额外的查询预算，即可选择相似度最高的替代模型。攻击者可以通过在替代模型训练中留出一个k^{th} 查询样本，并进行 k折交叉验证，来计算相似度。

6 评估

本节介绍数据集、评估设置与指标，以及获得的结果。

6.1 数据集

为了评估 AUTOLYCUS 的性能，我们采用了六个常用数据集，其中大多数来自 UCIMachine LearningRepository [25]。有关数据集的详情，请参阅表 2。

表 2：数据集属性。

数据集名称	样本数	特征数 (m)	类别数 (t)	数据类型
鸢尾花 [25]	150	4	3	连续型
作物 [20]	1700	7	17	连续型
乳腺癌 [25]	569	30	2	连续型
成人收入 [25]	48,842	14	2	混合型
托儿所 [25]	12,960	8	3	类别型
蘑菇 [25]	8,124	22	2	类别型

6.2 评估设置和指标

评估设置。我们使用 scikit-learn [40] Python 软件包对模型进行预处理、训练和评估。在训练模型之前，我们通过删除数据集中的所有缺失项，并使用标签编码器对类别型变量进行编码，对数据集应用预处理技术。每个数据集被划分为三个子集：(i) 训练数据集（75%）、(ii) 测试数据集（15%）和 (iii) 辅助集（10%）。训练数据集用于训练目标模型 M，测试数据集用于评估模型提取攻击的性能，而辅助集则用于创建辅助数据集 D_A。我们从辅助集中随机生成一个子集，其中每个类别包含 n 个样本，从而创建一个最小数据集 D_A。值得注意的是，尤其对于类别数量较少的模型，攻击者可以手动生成 D_A，

表3：与最先进水平的比较

数据集	最先进攻击				AUTOLYCUS				
	攻击名称	模型	$1-R_{test}$	查询	模型	$1-R_{test}$	查询	n	可解释人工智能工具
Iris	方程求解 [55]	逻辑回归	1	644	逻辑回归	1	**100**	1	LIME
Iris	路径查找 [8, 55]	决策树	1	246	决策树	1	**10**	1	LIME
Iris	IWAL [8]	决策树	1	361	决策树	1	**10**	1	LIME
乳腺癌	方程求解	逻辑回归	1	[644,1485]	逻辑回归	0.992	**100**	5	SHAP
成人收入	方程求解	逻辑回归	1	1485	逻辑回归	0.998	**1000**	5	SHAP
成人收入	路径查找	决策树	1	18323	决策树	0.937	**1000**	5	SHAP
成人收入	IWAL	决策树	1	244188	决策树	0.937	**1000**	5	SHAP

通过随机采样，且无需任何先验知识。这一观察结果与攻击者缺乏辅助数据的情形高度吻合。随后，攻击者使用 D_A 启动遍历算法（如第5节.2所述）。

我们在以下可解释模型上进行实验：决策树、逻辑回归、朴素贝叶斯、k 近邻分类器和随机森林。我们还测试 AUTOLYCUS 对多层感知机的适用性，并在第8节讨论其局限性。为减轻每次模型训练过程中可能产生的偏见，我们使用不同的随机状态在数据集 D_S 上训练 50 个代理模型，并生成 10 个辅助数据集 D_A 以初始化遍历。对于每个辅助数据集 D_A，我们将每项实验重复 10 次，并报告结果的平均值。

评估指标。 我们通过两种不同的指标，评估所构建的代理模型相较于目标模型的性能。具体而言，我们采用准确率和相似度指标，衡量所构建的代理模型在测试数据集上的效果。准确率表示代理模型所有预测中正确分类所占的比例。相比之下，模型相似度是目标模型与代理模型在中性数据集上的标签一致性；文献中也将其称为保真度或 $1-R_{test}$ [55]。在测试数据集 D上， R_{test}的形式化定义为 $R_{test}(f, \hat{f}) = \sum_{(\boldsymbol{x},\ y)\in D} d(f(\boldsymbol{x}), \hat{f}(\boldsymbol{x}))/|D|$。这使我们能够确定生成的代理模型在多大程度上能够复现目标模型的功能。**与其他技术的比较**。 我们将 AUTOLYCUS 的性能与另外四种方法进行比较。

基线攻击。 代理模型 S直接使用辅助数据集（$D_A = D_S$）进行训练。在这种情形下，攻击者不会向目标模型 M发送任何查询（$Q = 0$）。

***Steal-ML*攻击。** Tramer 等人提出了一种针对决策树模型的“路径搜索攻击”，以及一种针对逻辑回归模型的方程求解攻击，用于模型提取 [55]。路径搜索攻击[1] 是一种确定性的自顶向下攻击，会遍历所有节点，直至实现精确重建。在[55], 中，攻击者通过提取节点谓词并创建替代样本，利用树路径中的分裂。这里，节点谓词是样本必须满足的条件，满足后才能到达同一个节点ID。这种攻击的缺点是依赖较强的假设，例如节点ID 可直接获取，并且发送不完整的查询以便进行探索从树顶向树底进行探索。“方程求解攻击”是一种针对逻辑回归模型和神经网络的技术。该方法将查询结果转换为线性方程，并通过联立求解得出结果。根据具体设置，提取可能需要数十到数百次查询（针对神经网络时甚至可能需要数百万次）。该方法使用置信区间作为辅助模型信息，这会无意中泄露模型的内部机制，从而忽视了黑盒访问场景的背景。

***IWAL*攻击。** Chandrasekaran 等人 [8]提出了一种基于主动学习、针对树结构模型的模型提取攻击 IWAL。IWAL 是 Beygelzimer 等人提出的重要性加权主动学习算法 [5], ，它通过最小化标注错误，在每次查询中迭代地细化一棵树。

6.3 实验

我们评估攻击配置参数对 AUTOLYCUS 性能的影响：(i) 查询预算（Q），通常指可发送至机器学习即服务平台的最大查询数量；(ii) 攻击者的辅助数据集（D_A）中每个类别的样本数（n）；以及 (iii) 解释方法。更多详情见图4和图5。对于每种对应的模型类型和数据集，我们将相似度结果以及 AUTOLYCUS 所需的查询预算与基线攻击、Steal-ML 和 IWAL攻击所需的结果和预算进行比较。这些攻击的结果取自各自的论文。如果上述论文未明确记录或讨论某些数据集和模型类型，则仅在相关信息可用时进行比较。文献中还有其他值得考虑的重要攻击类型，例如 EAT [8], 、自适应再训练 [8], 、Lowd和Meek [31], 以及反事实 [1, 58] （见第8节）。然而，由于这些攻击所针对的模型类型（例如不可解释模型）与本研究不兼容，因此无法纳入本研究。有关我们在最先进攻击之间最佳结果比较的更多详情见表3。需要强调的是，在我们特定的攻击中，在当前假设下，LIME解释提供的信息显著更多。这种差异源于 SHAP 只提供特征重要性，却不为扰动提供方向指引。它是一种半盲查询方法，依据排名靠前的特征顺序并谨慎选择 δ来引导查询。使用 SHAP 的攻击只有在局部决策边界无法充分传达全局模型行为时，才会胜过使用 LIME 的攻击。

[1]路径搜索结果取自 Chandrasekaran 等人 [8]复现的结果，而非原始论文 [55]

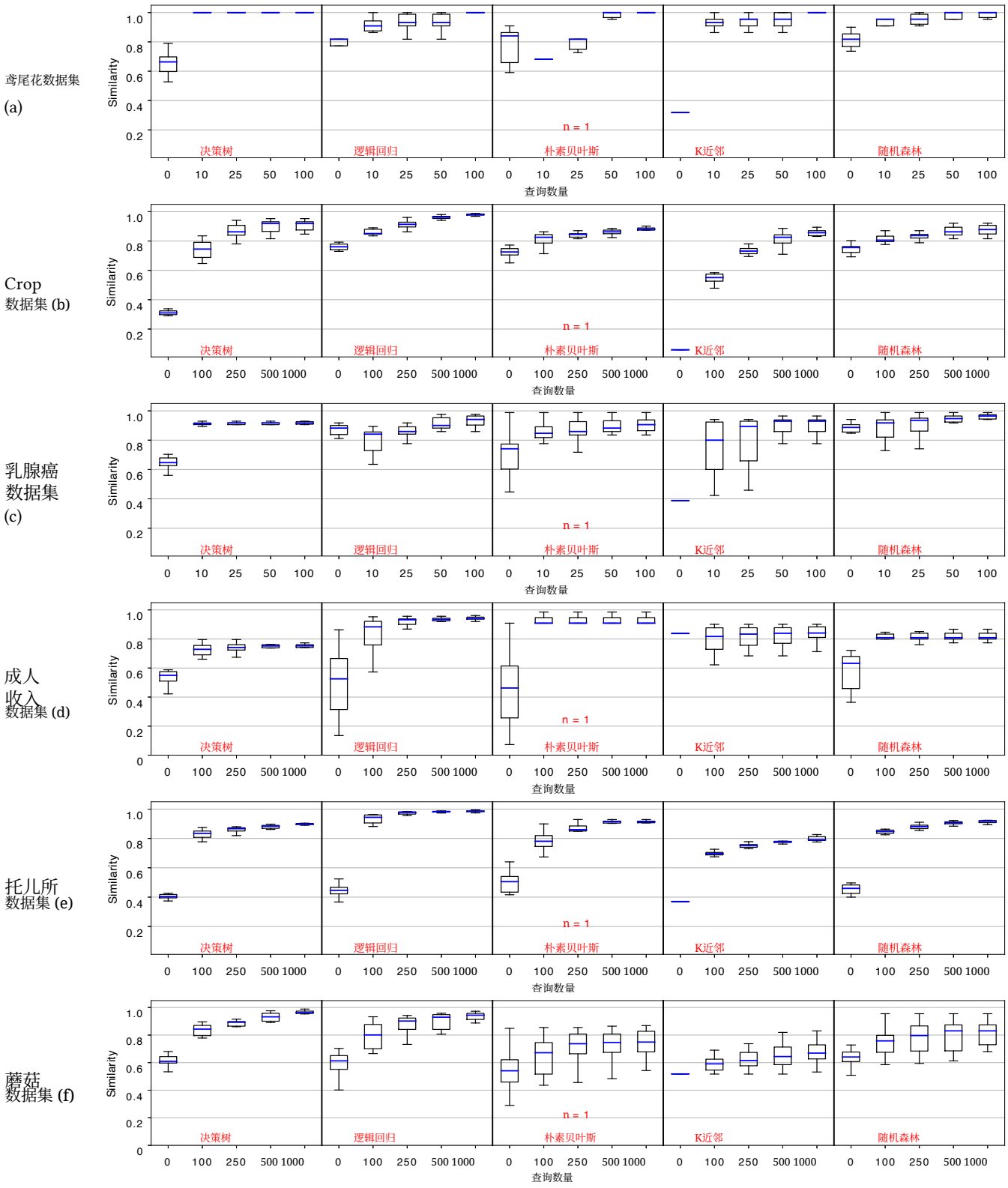


图4：查询数量的影响（Q）对使用 LIME 时代理模型相似度的影响。（$k = 3,\ \ n = 1$）

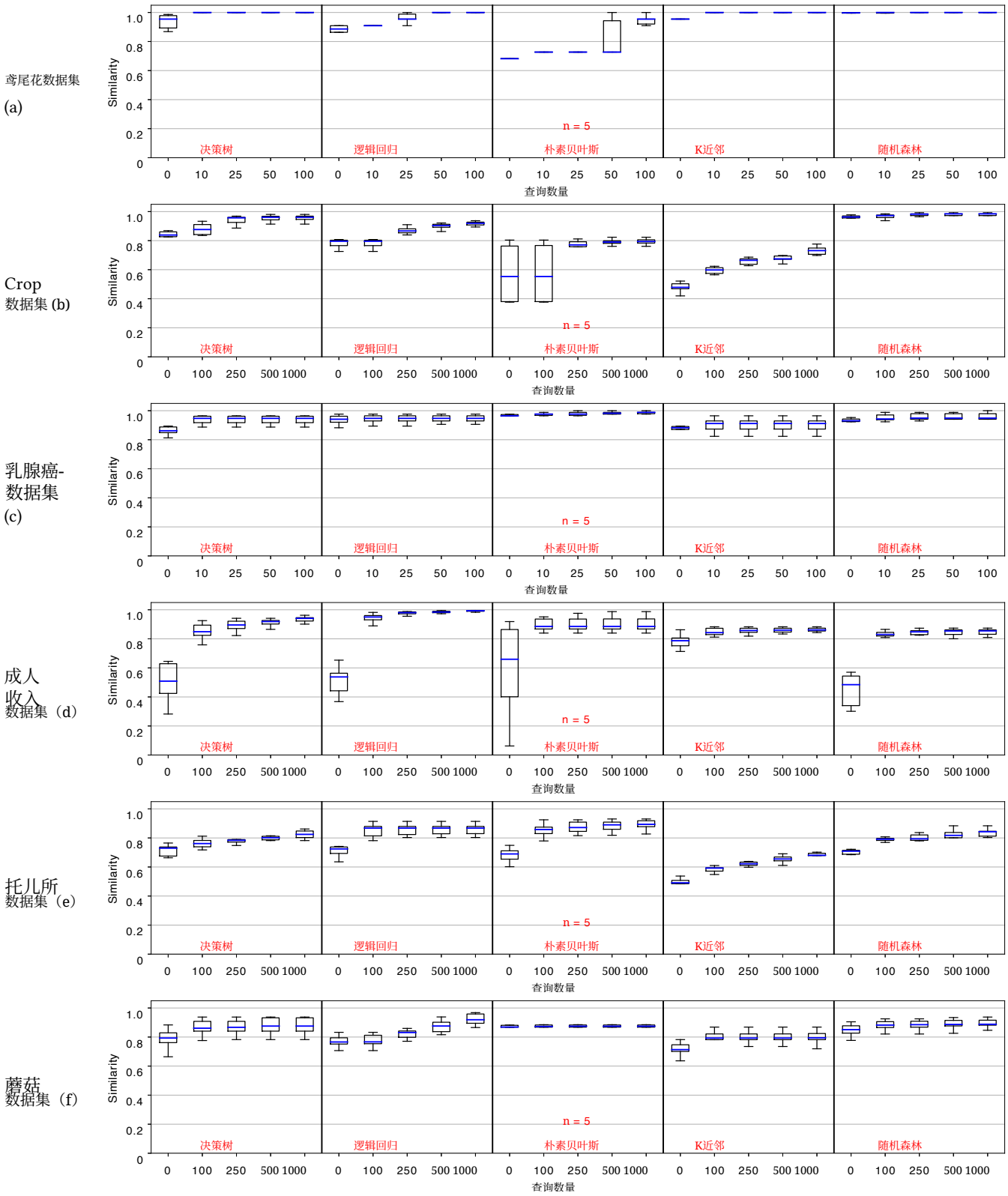


图5：查询数量的影响 (Q) **对使用 SHAP 时代理模型相似度的影响。**$(k = 3, n = 5$)

对于图4、图5和表3，允许探索的排名靠前的特征数量（k）设为 3。每个类别的辅助数据集大小（n）设为 1（用于 LIME）和 5（用于 SHAP）。这是出于设计考虑，旨在展示辅助数据集大小的影响，同时考虑到前述缺点，略微向 SHAP 倾斜。在采用类似 n设置的补充实验中（SHAP 为 $n = 1$，或 LIME 为 $n = 5$），由于信息量较少，SHAP 的表现不如 LIME。最后，SHAP 攻击的 δ取自辅助数据集中观测到的标准差和解空间划分。

鸢尾花数据集。对于所有考察的机器学习模型，我们观察到目标模型的准确率均为 1。攻击者只需向大多数目标模型发送 10 次查询（如图4和图5(a)所示），即可获得一个与目标模型 M 相似度接近 1 的代理模型 S。这是因为在鸢尾花数据集上训练的模型较为简单。由于 SHAP 的 n值更高，其表现优于 LIME。当 SHAP 中 $Q = 10$ 时，查询相似度接近基线攻击，这进一步证实了这一点。除朴素贝叶斯外，在 $Q < 50$时，AUTOLYCUS 在 LIME 和 SHAP 攻击中均显著优于基线方法。与表3中针对决策树和逻辑回归模型的 Steal-ML 和 IWAL 攻击所报告的查询数量相比，AUTOLYCUS 所需的查询更少，尤其是在辅助数据集增大时。

作物数据集。在我们考虑的数据集中，作物数据集包含最多的类别（t）。除朴素贝叶斯模型的准确率达到0.84外，目标模型的准确率均高于0.98。由于作物数据集中具有较高的 t，随着查询数量的增加，所有模型的模型相似度逐步提升的趋势更为明显。当 $Q = 1000$时，在图4(b)中，逻辑回归模型的相似度接近1。对于图5(b)，决策树和随机森林模型也具有这样的相似度。通过将这些结果与鸢尾花数据集所得结果进行比较，我们得出两点观察。首先，正如Rigaki等人 [43]所述，随着模型复杂度（架构、特征数量和类别数量）降低或模型准确率提高，模型提取攻击的性能得到增强。其次，即使使用额外的数据集，LIME有时仍因其解释的信息量较高而优于SHAP。

乳腺癌数据集。 对于乳腺癌数据集，目标模型的准确率介于 0.86 到 1 之间。在图5(c)中，我们观察到，只要已知足够多的真实样本，即使是基线攻击也能生成准确率足够高的模型。在图4(c)中，我们观察到，随着查询数量从 0 增至 100，不同模型类型的目标模型与替代模型之间的整体相似度持续提高。除决策树模型外，仅用 100 次查询，总能找到至少一个与目标模型相似度近乎完美的替代模型。观察 LIME 和 SHAP 攻击时，随着 n 从 1 增至 5，决策树模型的性能并未提升。这是因为解释所提供的决策边界未必覆盖目标模型的整个解空间，某些类别或分支可能仍未被探索。我们在该数据集上进行的其他实验表明，使用更高的 k 能够实现完美提取，这可以用重要特征的分布更加均匀来解释。尽管在该数据集上评估了方程求解攻击，但它并未由 Tramer 等人 [55] 提取在该数据集上训练的逻辑回归模型所需的查询数量。根据针对鸢尾花数据集和成人收入数据集的表现，前者通过 644 次查询提取出较简单的模型，后者则通过 1485 次查询提取出较复杂的模型；据此，我们可以推测，在乳腺癌数据集上训练的逻辑回归模型，所需查询数量可能介于两者之间。值得注意的是，AUTOLYCUS 的表现优于方程求解攻击，仅用 100 次查询便完成提取（查询数量减少了 6.4 倍）。

成人收入数据集。 对于图 4(d) 和 5(d) 中的成人收入数据集，目标模型的准确率约为 0.8。成人收入数据集包含 100 多个特征（编码后），并会生成树深度超过 20 的复杂树模型。这是一个关键数据集，凸显了优化所有攻击配置参数以获得最佳性能的重要性。它需要更高的查询限制、考虑更多特征、数量更多且种类多样的辅助样本，以及一组经过充分选择、用于扰动的 δ 值。这些依赖关系因模型类型而异。对于逻辑回归，使用 SHAP 攻击时，1000 次查询即可获得与目标模型相似度为 0.998 的代理模型，而方程求解攻击则需要 1485 次查询。对于决策树模型，在 SHAP 攻击中优化设置 $n = 5$ 和 δ时，1000 次查询即可获得相似度为 0.937 的模型；相比之下，路径查找攻击和 IWAL 攻击分别需要 18323 次和 244188 次查询。采用更优的配置参数，还能进一步提升性能。

托儿所数据集。 对于托儿所数据集，除朴素贝叶斯外，目标模型的准确率介于 0.95 到 1 之间。该数据集包含八个非二元（编码后超过 20 个类别）的分类特征。总体而言，不同模型之间的相似度在 LIME 攻击中约为 0.91（图4(e)），在 SHAP 攻击中约为 0.81（图5(e)）。在此数据集中，LIME 提取的结果始终优于 SHAP。这是因为特征重要性的分布较为均匀，而 LIME 凭借所提供的、更具信息量的决策边界克服了这一问题。当 k 增加到接近 m（特征总数）的取值时，即使 n 较低，SHAP 的表现也与 LIME 相近。

蘑菇数据集。 与托儿所数据集类似，蘑菇数据集的目标模型准确率介于 0.95 到 1 之间，但朴素贝叶斯的准确率为 0.8。蘑菇数据集与成人收入数据集的特征数量（编码后）最多，因此解空间相当庞大。在决策树和逻辑回归的情况下，LIME 攻击仅需 1000 次查询即可获得相似度接近或等于 1 的代理模型。然而，除逻辑回归外，与基线攻击相比，SHAP 攻击在各种模型类型上的相似度均停滞不前。在两种攻击中，朴素贝叶斯和 K 近邻模型提取的效果都比其他模型提升缓慢或停滞。这表明，与托儿所数据集类似，当 $k = 3$时，它未能涵盖所有重要特征，因此需要增加 k。此外，在 LIME 攻击中，对于固定的 n 取值，不同辅助数据集下攻击性能的差异很大。这表明，辅助数据的信息量可能会大幅降低或提升这些模型的提取性能。

探索的特征数量。 允许遍历的特征数量（k）对影响 AUTOLYCUS 的性能起着关键作用

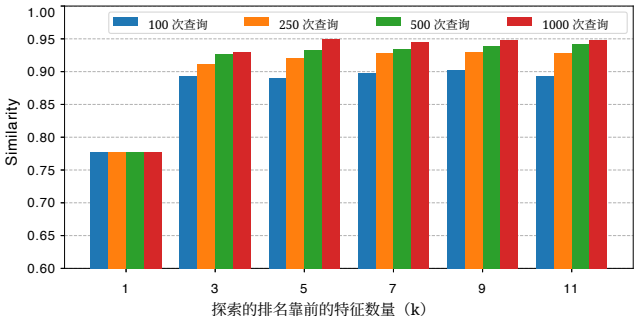


图 6： k **对代理模型相似度的影响。**（$n = 3$， $LIME$）

。在实验中，我们设置为 $k = 3$ （如图4和图5所示）。对于使用特征数量有限的数据集训练的模型，这一数值可能足够；但对于更大型的模型，尤其是有更多特征会影响模型决策时，这一数值可能不足。

我们展示了 k 对在成人收入数据集上训练的逻辑回归模型进行 LIME 攻击的影响。请注意， k 对 SHAP 攻击的影响与 LIME 攻击相似，因为这两种攻击都利用特征重要性。图 6 显示了在改变 k时，不同查询数量下代理模型与目标模型的相似度。对于 $k = 1$，探索主要围绕单个特征展开，因此代理模型的相似度最高只能达到 0.77。随着 k增加至 5，在生成 1000 次查询时，代理模型的相似度最高可达 0.95。超过 $k > 4$后，相似度的提升变得微乎其微，这表明目标模型依赖大约 5 到 6 个关键特征。

7 对策

本节探讨五种可用于防御 AUTOLYCUS 的对策。其中两种基于差分隐私 [14]。这些方法受到 Tramer 等人关于预测最小化 [55] 的建议启发。

实验中探索的一种对策是差分隐私，它可作为防御模型提取攻击的防御机制。我们考察了三种差分隐私方法。第一种方法是在目标模型的训练数据集 M中添加噪声。正如 Tramer 等人所指出的，这种方法无法缓解模型提取攻击，因为其目标并不一致；添加噪声可能只会妨碍对训练数据集敏感信息的学习，却没有针对模型的边界或参数。我们在训练数据中使用 ϵ 值介于 [1,ln(3)] 之间的校准拉普拉斯噪声，应用朴素差分隐私方法，以此验证这一点。我们发现，这种做法无法缓解模型提取攻击，也不会对决策边界产生显著影响。因此，正如 Tramer 等人所建议的，我们将工作重点转向针对模型边界。

我们探讨的第二种差分隐私方法是动态失真。该方法直接向 LIME 解释的决策边界添加噪声，并随着新的查询发送至模型，改变这些边界。在 LIME 中，每个特征的决策边界都是静态的有序列表，即每个模型对应 $DB^j = \{db^j 1, db^j 2, ...\}$，其中 db_i^j 等于 x_i^j 的邻居。[2] 为实现这一对策，我们首先确定 Q 个查询作为隐私预算，并从目标模型的训练数据集中移除 Q 个样本。这样，对于每个样本的决策区域，我们就能探究解释的决策边界在任意一对相邻模型 M 和 $M\prime$之间偏移了多少。请注意，这些模型分别使用 $n * t$ 个和 $n * t - Q$ 个样本进行训练。对于一个样本 X_i 以及所有相邻模型 M 和 $M\prime$，任意决策边界 d_i^j 可能产生的最大差异可形式化表示如下： $\sigma_i^j = max||M（d_i^j） - M\prime（d_i^j）||$。然后，我们计算一个拉普拉斯分布，其偏差对应于$\sigma$ 作为敏感度；对于不同的 ϵ 值，我们直接向解释的决策边界添加拉普拉斯噪声，使其在每次查询时都发生变化。$\epsilon = 1$ 时动态失真的结果如图7（a）所示。该方法不仅未能缓解模型提取攻击，反而通过允许探索更多样且信息更丰富的代理数据集，增强了攻击的力度。

第三种方法是静态失真。这是一种通用的随机噪声应用方法，采用动态失真方法并使用 $\epsilon = 1$。不过，在这种方法中，从拉普拉斯分布中抽取的噪声在不同查询之间保持不变，因此模型解释看起来就像是从另一个模型中得出的。静态失真的结果如图7 (b) 所示。我们发现，这种方法同样无法缓解模型提取攻击；与前一种方法一样，它反而通过提高最终替代模型的准确率和相似度，使攻击更加强大。不过，与第二种方法相比，这种提升微不足道。

施加噪声时，使用足够低的 ϵ 值可以扭曲（动态和静态的）解释边界，从而防止模型提取攻击。然而，在此类情况下，模型解释的效用受到的阻碍大于模型提取的成功率。使用更低的 ϵ 值进行实验，由于过度扰动，产生了数学上不可能的决策边界。

第四种方法涉及向 SHAP 的特征重要性添加噪声。这些重要性根据攻击者定义的 δ_j 值，决定特征扰动的顺序。虽然显著的噪声可能会扰乱这一顺序，并导致接近 k^{th} 的特征扰动发生偏差，但除非 k 的值足够低，否则随着查询次数增加，这些偏差会趋于正常化。然而，这种正常化会损害解释的效用，可能误导用户对其预测的理解。因此，在 SHAP 解释中使用噪声不太可能有效抵御模型提取攻击。

Tramer 等人提出的另一种对策，是对预测过程中获得的置信度值进行舍入。这种方法直观上能够有效防御依赖置信度值、且通过确定性方式求解边界的模型提取攻击（即方程求解攻击、路径查找攻击）。由于 AUTOLYCUS 在重新训练过程中以随机方式反映决策边界，并且只使用标签和解释中的决策边界，

[2]特征状态（x_i^j值）之间的邻接关系基于相应的决策边界。

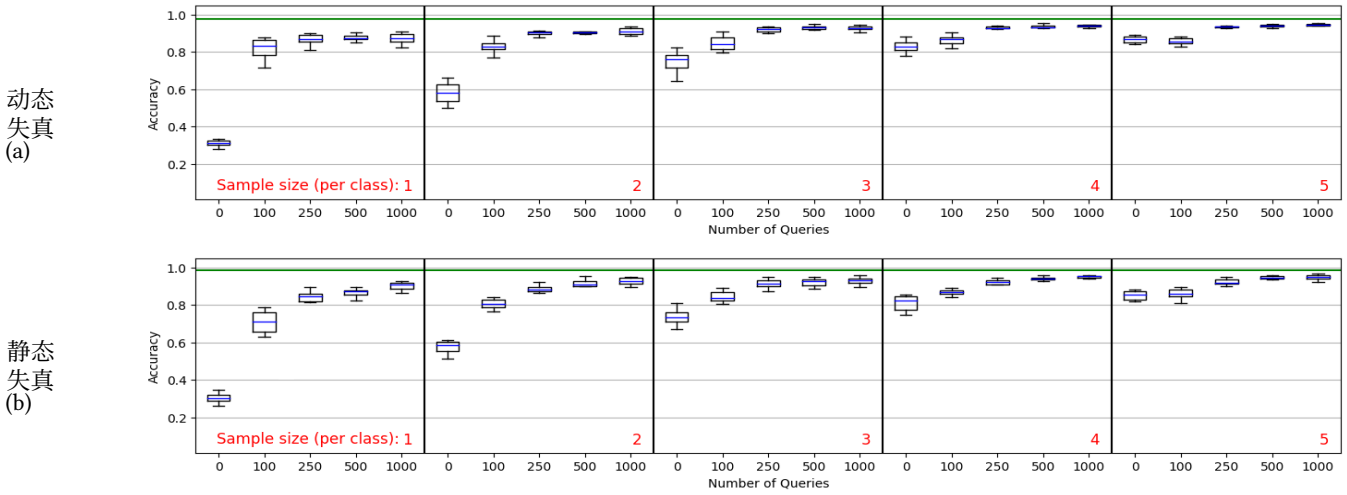


图7：对策对 Crop 数据集的影响。

因此，置信度舍入或调整解释强度等方法不会对其性能产生任何影响。

预测最小化方法在 AUTOLYCUS 中容易失效，但设置查询预算等对策可以减缓提取过程。尽管本研究尚未对此进行探索，针对恶意代理模型共享的指纹识别 [32] 是未来潜在的研究方向。

8 讨论

局限性。 AUTOLYCUS 的主要局限性在于，与 Tramer 等人 [55]提出的白盒攻击（方程求解、路径搜索）所实现的精确重构相比，它只能以概率方式重构模型。AUTOLYCUS 需要在黑盒设置下重新训练模型，因此，解释、遍历或辅助数据集等攻击配置参数在某些模型中会受到限制。

例如，未能在解释或辅助数据集中得到充分表示的异常值分类（例如决策树模型中的边界情况）可能会在遍历过程中被遗漏。在具有大量排名靠前的特征的复杂模型中，这一局限性尤为明显，因为某些分支可能无法在代理模型中得到恰当表示，从而降低其与目标模型的相似性。在使用 LIME 针对在成人收入数据集上训练的决策树模型进行攻击的实验中，可以观察到一个局限性示例。尽管生成的代理模型具有相近的准确率，但该模型的准确率较低且包含多个重要特征，使得在现有假设和规模较小的辅助数据集条件下，难以充分发现并以功能方式表示模型的某些方面。

此外，AUTOLYCUS 的有效性仅限于目标模型所使用的训练数据类型。AUTOLYCUS 在适用于可解释模型和决策边界的表格数据及混合数据上表现良好。然而，在图像分类任务中，针对单个像素的表格型 LIME 和 SHAP 解释具有相同的强度，因此将这些特征和边界单独考察并无助益，实际上使攻击变得无用。由于图像分类会考察图像中的像素或区域集合，因此必须采用适用于此类任务的解释和遍历方法。因此，AUTOLYCUS 应用于具有信息丰富的解释的模型。

与第6.3节中的评估不同，AUTOLYCUS 的性能会受到以下攻击配置和模型参数精细调整的影响：(i) 确保发送足够多的查询；(ii) 根据特征重要性的分布，选择适当数量的特征（k）进行扰动；(iii) 纳入足够数量的 (iv) 信息丰富的辅助样本；(v) 考虑模型的复杂度——复杂度较低的模型更容易被提取——包括神经网络术语中任何未知项或参数；(vi) 确保模型的准确率；(vii) 优化 δ 值的选择，尤其是在使用 SHAP 作为可解释人工智能工具时。这些因素共同提升了 AUTOLYCUS 在对抗环境中的效能。

其他模型无关的可解释性工具。 在其他解释技术中，加入反事实可以提供额外的宝贵信息，文献中也记录了各种相关攻击 [1, 58]。我们使用 LIME 和 SHAP 解释的目标，与通过额外步骤复现反事实的效用相似。虽然直接提供反事实可以简化我们的框架，但也会使其创新性有限。此外，由于已有关于隐私风险的记录，机器学习即服务提供商可能会避免在商业环境中使用这些解释。出于这些考虑，我们选择不在框架中明确纳入反事实。

应用于其他机器学习模型。在针对可解释模型取得了令人鼓舞的结果后，我们研究了 AUTOLYCUS 对黑盒模型的有效性和可迁移性。我们首先关注神经网络。我们在四个数据集上进行了实验，配置与 Steal-ML 攻击类似：目标模型由多层感知机组成，且仅包含一个隐藏层，该层包含 20 个节点。由于乳腺癌数据集上的模型在包含 20 个节点时准确率仅为 0。22，因此仅针对该数据集将隐藏层节点数增加至 100。我们设置 $k = 3$

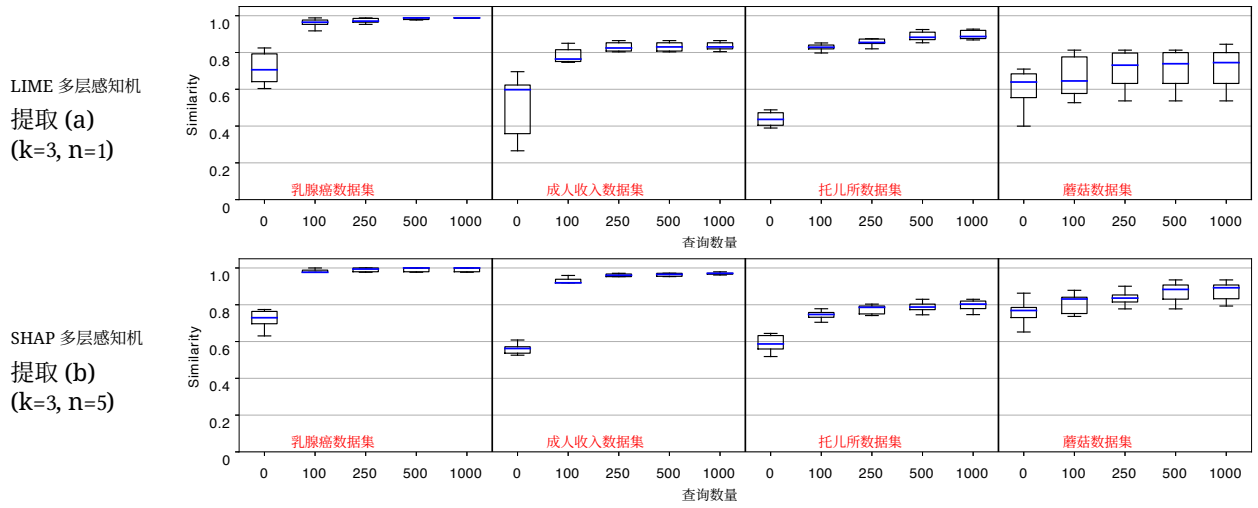


图 8：不同数据集上的多层感知机代理模型相似度。

并设置 $n = 1$（针对 LIME）和 $n = 5$（针对 SHAP），遵循第 6.3 节中的设置。图 8 展示了各数据集的相似度。

对于乳腺癌数据集，攻击者在 LIME 攻击和 SHAP 攻击中均仅用 100 次查询，就获得了相似度最高达 1 的代理模型。鉴于该数据集的复杂度相对较低，这一结果符合我们的预期；图 4 和图 5(c) 所示的结果也印证了这一点。此外，对于成人收入和托儿所数据集，当目标模型为随机森林且使用 LIME 时，代理模型的性能相近。另一方面，尽管辅助数据集规模扩大使 SHAP 攻击优于 LIME 攻击，但与 LIME 攻击相比，它在提取基于托儿所数据集训练的神经网络时效率较低。这种差异可归因于解释信息量的不同。当 k 较低时，LIME 擅长解释未知项较少的分类解空间。与成人收入和蘑菇数据集相比，托儿所数据集的解空间相对较小，因此即使 n 的值较低，LIME解释仍比 SHAP 解释更有效。然而，随着 k增加，SHAP 攻击的性能开始提升，这与 LIME 解释所呈现的情况类似。在蘑菇数据集中，增加查询数量并未使 LIME攻击的代理模型相似度得到预期提升。将 k 设为较高值的实验也未能带来在可解释模型中观察到的显著性能提升。这一局限可能是由于查询数量不足，无法准确表示神经网络的复杂边界；此外，解空间较大的特征数量较多，也使得在查询预算相对较低时，提取过程更加困难。在 SHAP 攻击中，由于可用辅助样本数量较多（n），提取更为成功，而且查询数量越多，效果通常越好。这表明，样本越充足，就越有利于有效探索较罕见的情形（这类情形在蘑菇数据集中似乎更多）。此外，在提取神经网络的复杂边界时，LIME 和 SHAP 解释可能偶尔会出现不一致，这也可能成为影响性能的另一个因素。总之，在条件相当且 n 和 k的值足够大时，与 SHAP 相比，LIME 在捕捉神经网络中的相似性方面通常表现更优。

9 结论

本文提出了 AUTOLYCUS，这是一种新颖的模型无关提取攻击，利用模型无关的人工智能解释来提取可解释模型的决策边界。在六个不同数据集上的实证结果表明，该攻击具有很高的有效性。具体而言，我们证明了攻击者可以利用人工智能解释创建相当准确的替代模型；即使查询预算较低，这些替代模型仍与目标模型具有很高的相似性。我们还证明，与依赖精确重建的最先进的攻击相比，AUTOLYCUS 在准确率和相似性相当的情况下，仅需更少的查询即可实现部分重建。此外，我们还探讨了减轻此类攻击的潜在对策，例如在解释的决策边界中加入噪声，或在训练数据中加入噪声。实证结果表明，这些对策对 AUTOLYCUS 并不起作用。未来，我们计划研究多种模型解释的有效性、黑盒机器学习模型（如卷积神经网络）中因解释而增加的潜在风险，以及新的有效对策。

致谢

对于本出版物所报告的研究，Abdullah Caglar Ok- suz 和 Erman Ayday 获得了美国国家科学基金会（NSF）项目 OAC-2112606 的部分资助。Anisa Hal- imi 部分获得了欧盟地平线2020研究与创新计划项目 951911 – AI4Media 的资助。本文内容完全由作者负责，并不一定代表资助本研究的机构的官方观点。

参考文献

[1] Ulrich Aïvodji、Alexandre Bolot 和 Sébastien Gambs。2020。基于反事实解释的模型提取。arXiv:2009.01884 https://arxiv.org/abs/2009. 01884[2] Microsoft Azure。2024。模型可解释性 - Azure Machine Learn- ing。https://learn.microsoft.com/en-us/azure/machine-learning/how-to-machine-learning-interpretability?view=azureml-api-2 [访问日期：01-Mar- 2024].[3] Sebastian Bach、Alexander Binder、Grégoire Montavon、Frederick Klauschen、Klaus-Robert Müller 和 Wojciech Samek。2015。通过逐层相关性传播对非线性分类器决策进行像素级解释。*PloS one* 10, 7 2015 0130140 10 1371 ( ), e 。https://doi.org/ 。/ journal.pone.[4] Alejandro Barredo Arrieta、Natalia Díaz-Rodríguez、Javier Del Ser、Adrien Ben- netot、Siha m Tabik、Alberto Barbado、Salvador Garcia、Sergio Gil-Lopez、Daniel Molina、Richard Benjamins、Raja Chatila 和 Francisco Herrera。2020。可解释人工智能（XAI）：面向负责任人工智能的概念、分类体系、机遇与挑战。*Information Fusion* 58（2020），82–115。https://doi.org/ 10 1016 2019 12 012 。/j.inffus。。Alina Beygelzimer、Daniel Hsu、John Langford 和 Tong Zhang。2010。无约束的无偏主动学习。载于 神经信息处理系统第*23*届国际会议论文集 - 第*1*卷 （加拿大不列颠哥伦比亚省温哥华）（*NIPS'10*）。Curran Associates Inc.，美国纽约州雷德胡克，199–207。https://proceedings.neurips.cc/paper/2010/hash/00411460f7c92d2124a67ea0f4cb5f85-Abstract.html[6] Nicholas Carlini、Chang Liu、Úlfar Erlingsson、Jernej Kos 和 Dawn Song。2019。《秘密共享者》：评估和测试神经网络中的非预期记忆。载于 第*28*届 *USENIX* 安全研讨会、*USENIX Security 2019*，美国加利福尼亚州圣克拉拉，*2019*年*8*月*14*日至*16*日，Nadia Heninger 和 Patrick Traynor（编）。USENIX 协会，美国加利福尼亚州圣克拉拉，267–284。https://www.usenix.org/conference/usenixsecurity19/presentation/carlini [7] IBM Thomas J Watson 研究中心。2019。可解释人工智能（XAI）。https://www.ibm.com/watson/explainable-ai[8] Varun Chandrasekaran、Kamalika Chaudhuri、Irene Giacomelli、Somesh Jha 和 Songbai Yan。2020。探索主动学习与模型提取之间的联系。载于 第*29*届 *USENIX* 安全研讨会、*USENIX Security 2020*，*2020*年*8*月*12*日至*14*日，Srdjan Capkun 和 Franziska Roesner（编）。USENIX 协会，美国，1309–1326。https://www.usenix.org/conference/usenixsecurity20/presentation/ chandrasekaran[9] Christopher A. Choquette-Choo、Florian Tramèr、Nicholas Carlini 和 Nicolas Papernot。2021。仅标签成员推断攻击。载于 第*38*届国际机器学习会议论文集，*ICML 2021*，*2021*年7月*18*日至*24*日，线上活动（机器学习研究论文集，第*139*卷），Marina Meila 和 Tong Zhang（编）。PMLR，美国，1964–1974。http://proceedings.mlr.press/v139/choquette-choo21a.html [10] Google Cloud。2021。Vertex 可解释人工智能。https://cloud.google.com/vertex- ai/docs/explainable-ai/overview [访问日期：24-Feb-2024]。[11] IBM Cloud。2023。Watson Openscale。https://www.ibm.com/docs/en/cloud- paks/cp-data/4.8.x?topic=models-configuring-model-evaluations [访问日期：24- Feb-2024]。[12] Filip Karlo Dosilovic、Mario Brcic 和 Nikica Hlupic。2018。可解释人工智能：综述。载于 第*41*届信息与通信技术、电子与微电子国际大会，*MIPRO 2018*，克罗地亚奥帕蒂亚，*2018*年*5*月*21*日至*25*日，Karolj Skala、Marko Koricic、Tihana Galinac Grbac、Marina Cicin-Sain、Vlado Sruk、Slobodan Ribaric、Stjepan Gros、Boris Vrdoljak、Mladen Mauher、Edvard Tijan、Predrag Pale 和 Matej Janjic（编）。IEEE，美国，210 215 10 23919 2018 8400040 – 。https://doi.org/ 。/MIPRO。 。[13] Vasisht Duddu、Debasis Samanta、D. Vijay Rao 和 Valentina Emilia Balas。2018。通过时序侧信道窃取神经网络。arXiv:1812.11720 http://arxiv.org/abs/1812 11720 。Cynthia Dwork、Frank McSherry、Kobbi Nissim 和[14] Adam D. Smith。2006。将噪声校准至隐私数据分析中的敏感度。载于 密码学理论、第三届密码学理论会议、*TCC 2006*，美国纽约州纽约，*2006*年*3*月*4*日至*7*日，论文集（计算机科学讲义，第*3876*卷），Shai Halevi 和 Tal Rabin（编）。Springer，美国，265 284 10 1007 11681878_14 – 。2019 https://doi.org/ 。/ Aaron Fisher、Cynthia Rudin 和 Francesca Dominici。 。所有模型都是错误的，但许多模型是有用的：通过同时研究整个预测模型类别来学习变量的重要性。*J. Mach. Learn. Res.* 20（2019），177:1–177:81。http://jmlr.org/papers/v20/18-760.html [16] Matt Fredrikson、Somesh Jha 和 Thomas Ristenpart。2015。利用置信度信息的模型反演攻击及基本对策。载于 第*22*届 *ACM SIGSAC* 计算机与通信安全会议论文集（美国科罗拉多州丹佛）（*CCS '15*）。计算机协会，美国纽约州纽约，1322 1333 10 1145 2810103 2813677 – 。[17] https://doi.org/ 。/ 。Jerome H. Friedman。 。贪心函数逼近：梯度提升机。统计年鉴 29, 5（2001），1189–1232。https://doi.org/10. 1214/aos/1013203451 2001

[18] Karan Ganju、Qi Wang、Wei Yang、Carl A. Gunter 和 Nikita Borisov。2018。利用置换不变表示对全连接神经网络实施属性推断攻击。载于*2018*年 *ACM SIGSAC* 计算机与通信安全会议论文集 （加拿大多伦多） （*CCS '18*）。计算机协会，纽约州纽约，美国，619–633。https://doi.org/10.1145/3243734.3243834[19] Andreas Holzinger、Anna Saranti、Christoph Molnar、Przemyslaw Biecek 和 Wojciech Samek。2020。可解释人工智能方法——简要概述。载于*xxAI*——超越可解释人工智能——与 *ICML 2020* 联合举办的国际研讨会，*2020*年*7*月*18*日，奥地利维也纳，修订和扩展论文（计算机科学讲义，第 *13200* 卷），Andreas Holzinger、Randy Goebel、Ruth Fong、Taesup Moon、Klaus-Robert Müller 和 Wojciech Samek（编）。Springer，美国， 13 38 10 1007 978 3 031 04083 2_2 – 。https://doi.org/ 。 / - - - - Atharva Ingle 、 N Nithin Srivatsav 和 Swathi N Shayana。2020。作物推荐数据集。https://www.kaggle.com/datasets/atharvaingle/crop-recommendation-dataset[21] Matthew Jagielski、Nicholas Carlini、David Berthelot、Alex Kurakin 和 Nicolas Papernot。2020。高准确率、高保真度的神经网络提取。载于第*29*届 *USENIX* 安全研讨会（*USENIX Security 2020*），*2020*年*8*月*12*日至*14*日，Srdjan Capkun 和 Franziska Roesner（编）。USENIX 协会，美国，1345–1362。https://www.usenix.org/conference/usenixsecurity20/presentation/jagielski [22] Mika Juuti、Sebastian Szyller、Samuel Marchal 和 N. Asokan。2019。PRADA：防御深度神经网络模型窃取攻击。载于*IEEE*欧洲安全与隐私研讨会（*EuroS&P 2019*），瑞典斯德哥尔摩，*2019*年*6*月*17*日至*19*日。IEEE，美国，512 527 10 1109 2019 00044 – 。https://doi.org/ 。 /[23] EUROSP. 。Andrei Kapishnikov、Tolga Bolukbasi、Fernanda B. Viégas 和 Michael Terry。2019。XRAI：通过区域实现更好的归因。载于*2019*年 *IEEE/CVF*国际计算机视觉会议（*ICCV 2019*），韩国首尔，*2019*年*10*月*27*日至*11*月*2*日。IEEE，美国，4947–4956。https://doi.org/10.1109/ICCV.2019. 00505[24] Sanjay Kariyappa 和 Moinuddin K. Qureshi。2020。利用自适应误导信息防御模型窃取攻击。载于*2020*年 *IEEE/CVF* 计算机视觉与模式识别会议（*CVPR 2020*），美国华盛顿州西雅图，*2020*年*6*月*13*日至*19*日。计算机视觉基金会 / IEEE，美国，767–775。https://doi.org/10.1109/CVPR42600.2020.00085[25] Markelle Kelly、Rachel Longjohn 和 Kolby Nottingham。2023。UCI机器学习知识库。https://archive.ics.uci.edu[26] Been Kim、Martin Wattenberg、Justin Gilmer、Carrie Cai、James Wexler、Fernanda Viegas 和 Rory sayres。2018。超越特征归因的可解释性：利用概念激活向量（TCAV）进行定量测试。载于第*35*届国际机器学习会议论文集（机器学习研究论文集，第*80*卷），Jennifer Dy 和 Andreas Krause（编）。PMLR，美国，2668–2677。https://proceedings.mlr.press/v80/kim18d.html [27] Pieter-Jan Kindermans、Kristof T. Schütt、Maximilian Alber、Klaus-Robert Müller、Dumitru Erhan、Been Kim 和 Sven Dähne。2018。学习如何解释神经网络：PatternNet 和 PatternAttribution。载于第*6*届国际学习表征会议（*ICLR 2018*），*2018*年*3*月*30*日，*4*月至*5*月，加拿大不列颠哥伦比亚省温哥华，会议专题论文集。OpenReview.net，美国。https://openreview.net/forum?id=Hkn7CBaTW[28] Alex Krizhevsky、Vinod Nair 和 Geoffrey Hinton。2010。CIFAR-10（加拿大高等研究院）。http://www.cs.toronto.edu/~kriz/cifar.html [29] Yann LeCun 和 Corinna Cortes。2010。MNIST手写数字数据库。http://yann.lecun.com/exdb/mnist/。http://yann.lecun.com/exdb/mnist/ [30] Zheng Li 和 Yang Zhang。2021。仅标签暴露下的成员关系泄露。载于*2021*年 *ACM SIGSAC* 计算机与通信安全会议论文集 （线上活动，大韩民国）（*CCS '21*）。计算机协会，纽约州纽约，美国，880–895。https: // doi.org/ 10 1145 3460120 3484575 . / 。Daniel Lowd 和 Christopher Meek。2005。对抗学习。载于第*11*届 *ACM SIGKDD*知识发现与数据挖掘国际会议论文集，美国伊利诺伊州芝加哥，*2005*年*8*月*21*日至*24*日，Robert Grossman、Roberto J. Bayardo 和 Kristin P. Bennett（编）。ACM，美国，641–647。https: //doi.org/ 10 1145 1081870 1081950 . / 。Nils Lukas、Yuxuan Zhang 和 Florian[31] Kerschbaum。2021。利用可转移对抗样本进行深度神经网络指纹识别。载于第*9*届国际学习表征会议（*ICLR 2021*），线上活动，奥地利，*2021*年*5*月*3*日至*7*日。OpenReview.net，美国。https://openreview.net/forum?id = 1 VqzVhqxkjH[32] Scott M. Lundberg 和 Su-In Lee。 。一种统一的模型预测解释方法。载于第*31*届神经信息处理系统国际会议论文集 （美国加利福尼亚州长滩） （*NIPS'17*）。Curran Associates Inc.，纽约州雷德胡克，美国，4768–4777。https://dl.acm.org/doi/pdf/10. 5555/3295222.3295230[34] Smitha Milli、Ludwig Schmidt、Anca D. Dragan 和 Moritz Hardt。2019。根据模型型解释重建模型。载于公平、问责与透明度会议论文集 （美国佐治亚州亚特兰大） （*FAT* '19*）。计算机协会，纽约州纽约，美国，1–9。https://doi.org/10. 1145/3287560.3287562 2017

[35] Takayuki Miura、Satoshi Hasegawa 和 Toshiki Shibahara。2021。MEGEX：针对基于梯度的可解释人工智能的无数据模型提取攻击。arXiv: 2107 08909 2107 08909 。https://arxiv.org/abs/ 。[36] Tribhuvanesh Orekondy、Bernt Schiele 和 Mario Fritz。2019。Knockoff Nets：窃取黑盒模型的功能。载于*IEEE*计算机视觉与模式识别会议（*CVPR 2019*），美国加利福尼亚州长滩，*2019*年*6*月*16*日至*20*日。计算机视觉基金会 / IEEE，美国，4954–4963。https://doi.org/10.1109/ CVPR.2019.00509[37] Guillermo Owen。2013。博弈论在经济学中的应用。*IGTR* 15, 3 (2013 1340011 10 1142 0219198913400112), 。[38] https://doi.org/ 。 /S Nicolas Papernot、Patrick McDaniel、Ian Goodfellow、Somesh Jha、Z. Berkay Celik 和 Ananthram Swami。2017。针对机器学习的实用黑盒攻击。载于*2017*年 *ACM* 亚洲计算机与通信安全会议论文集 （阿联酋阿布扎比）（*ASIA CCS '17*）。计算机协会，纽约州纽约，美国，506–519。https: //doi.org/ 10 1145 3052973 3053009 . / 。Nicolas Papernot、Patrick D. McDaniel 和 Ian J. Goodfellow。2016。机器学习中的可转移性：从现象到利用对抗样本实施黑盒攻击。arXiv: 1605 07277 1605 07277 。[40] http://arxiv.org/abs/ 。F. Pedregosa、G. Varoquaux、A. Gramfort、V. Michel、B. Thirion、O. Grisel、M. Blondel、P. Prettenhofer、R. Weiss、V. Dubourg、J. Vanderplas、A. Passos、D. Cournapeau、M. Brucher、M. Perrot 和 E. Duchesnay。2011。Scikit-learn：Python 机器学习。机器学习研究期刊 12（2011），2825–2830。https://www.jmlr.org/papers/v12/pedregosa11a.html [41] Marco Tulio Ribeiro、Sameer Singh 和 Carlos Guestrin。2016。"我为什么要相信你？ "：解释任意分类器的预测。载于第*22*届 *ACM SIGKDD*知识发现与数据挖掘国际会议论文集 （美国加利福尼亚州旧金山）（*KDD '16*）。计算机协会，纽约州纽约，美国，1135–1144。https://doi.org/10.1145/2939672. 2939778[42] Marco Túlio Ribeiro、Sameer Singh 和 Carlos Guestrin。2018。Anchors：高精度的模型无关解释。载于第三十二届 *AAAI* 人工智能会议（*AAAI-18*）、第*30*届创新人工智能应用会议（*IAAI-18*）及第*8*届 *AAAI* 人工智能教育进展研讨会（*EAAI-18*）论文集，美国路易斯安那州新奥尔良，*2018*年*2*月*2*日至*7*日，Sheila A. McIlraith 和 Kilian Q. Weinberger（编）。AAAI Press，美国，1527 1535 10 1609 32 1 11491 – 。https://doi.org/ 。 /[43] AAAI.V I 。Maria Rigaki 和 Sebastian García。。机器学习隐私攻击综述。*ACM* 计算调查 56, 4 （ 2024 ），101:1–101:34。https://doi.org/10.1145/ 3624010[44] Cynthia Rudin。2019。停止解释用于高风险决策的黑盒机器学习模型，改用可解释模型。*Nature Machine Intelligence* 1, 5（2019），206–215。https://doi.org/10.1038/s42256-019-0048-x[45] Ahmed Salem、Yang Zhang、Mathias Humbert、Pascal Berrang、Mario Fritz 和 Michael Backes。2019。ML-Leaks：针对机器学习模型的模型与数据无关的成员推断攻击及防御。载于第*26*届网络与分布式系统安全年度研讨会（*NDSS 2019*），美国加利福尼亚州圣迭戈，*2019*年*2*月*24*日至*27*日。互联网协会，美国。http://dx.doi. org/ 10 14722 2019 23119 . /NDSS. 。Ramprasaath R. Selvaraju、Michael Cogswell、Abhishek Das、Ramakrishna Vedantam、Devi Parikh 和 Dhruv Batra。2017。Grad-CAM：通过基于梯度的定位，从深度网络生成视觉解释。载于*IEEE*国际计算机视觉会议（*ICCV 2017*），意大利威尼斯，*2017*年*10*月*22*日至*29*日。IEEE 计算机学会，美国，618 626 10 1109 2017 74 – 。https://doi.org/ 。 /[47] ICCV. 。亚马逊云科技。 。Amazon SageMaker Clarify。https://aws.amazon.com/sagemaker/clarify/ [访问日期：2月24日-2024]。[48] Lloyd S. Shapley。1951。多人博弈札记——*I*：四人博弈的特征点解。兰德公司，美国加利福尼亚州圣莫尼卡。https: // doi.org/ 10 7249 0656 . /RM Yi Shi、Yalin Sagduyu 和 Alexander Grushin。2017。如何利用深度学习窃取机器学习分类器。载于*2017*年 *IEEE* 国土安全技术国际研讨会（*HST*）。IEEE，美国，1–5。https://doi.org/10. 1109/THS.2017.7943475[50] Reza Shokri、Martin Strobel 和 Yair Zick。2021。模型解释中的隐私风险。载于*AIES '21*：*AAAI/ACM* 人工智能、伦理与社会会议，线上活动，美国，*2021*年*5*月*19*日至*21*日，Marion Fourcade、Benjamin Kuipers、Seth Lazar 和 Deirdre K. Mulligan（编）。ACM，美国，231–241。https://doi.org/10.1145/ 3461702.3462533[51] Reza Shokri、Marco Stronati、Congzheng Song 和 Vitaly Shmatikov。2017。针对机器学习模型的成员推断攻击。载于*2017*年 *IEEE* 安全与隐私研讨会（*SP 2017*），美国加利福尼亚州圣何塞，*2017*年*5*月*22*日至*26*日。IEEE 计算机学会，美国，3 18 10 1109 2017 41 – 。https://doi.org/ 。[52] /SP. 。Karen Simonyan 、 Andrea Vedaldi 和 Andrew Zisserman。 。深入卷积网络内部：图像分类模型与显著图的可视化。arXiv: 1312 6034 [ ] 。cs.CV Congzheng Song 和 Vitaly Shmatikov。2020。过度学习揭示敏感属性。载于第*8*届国际学习表征会议（*ICLR 2020*），埃塞俄比亚亚的斯亚贝巴，*2020*年*4*月*26*日至*30*日。OpenReview.net，美国。https: //openreview.net/forum?id=SJeNz04tDS [54] Mukund Sundararajan、Ankur Taly 和 Qiqi Yan。2017。深度网络的公理归因。载于第*34*届国际机器学习会议论文集（*ICML 2017*），澳大利亚新南威尔士州悉尼，*2017*年*8*月*6*日至*11*日（机器学习研究论文集，第*70*卷），Doina Precup 和 Yee Whye Teh（编）。PMLR，美国，3319–3328。http://proceedings.mlr.press/v70/sundararajan17a.html [55] Florian Tramèr、Fan Zhang、Ari Juels、Michael K. Reiter 和 Thomas Ristenpart。2016。通过预测 API 窃取机器学习模型。载于第*25*届 *USENIX* 安全研讨会论文集（美国德克萨斯州奥斯汀）（*SEC'16*）。USENIX 协会，美国，601–618。https://dl.acm.org/doi/10.5555/3241094. 3241142[56] Jean-Baptiste Truong、Pratyush Maini、Robert J. Walls 和 Nicolas Papernot。2021。无数据模型提取。载于*IEEE* 计算机视觉与模式识别会议，*CVPR 2021*，线上，*2021*年*6*月*19*日至*25*日。计算机视觉基金会 / IEEE，美国，4771 4780 10 1109 46437 2021 00474 – 。[57] https://doi.org/ . /CVPR . . Binghui Wang 和 Neil Zhenqiang Gong。 。窃取机器学习中的超参数。载于*2018*年 *IEEE* 安全与隐私研讨会，*SP 2018*，论文集，*2018*年*5*月*21*日至*23*日，美国加利福尼亚州旧金山。IEEE 计算机学会，美国，36 52 10 1109 2018 00038 – 。https://doi.org/ . /SP. . Yongjie Wang、Hangwei Qian 和 Chunyan Miao。。DualCF：从反事实解释中高效提取模型的攻击。载于*FAccT '22*：*2022*年 *ACM* 公平性、问责制与透明度会议，韩国首尔，*2022*年*6*月*21*日至*24*日。ACM，美国，1318–1329。https://doi.org/10.1145/3531146. 3533188[59] Anli Yan、Teng Huang、Lishan Ke、Xiaozhang Liu、Qi Chen 和 Changyu Dong。2023。解释泄漏：解释引导的模型提取攻击。信息科学 632 2023 269 284 10 1016 2023 03 020 ( ), – 。https://doi.org/ . [60] /j.ins. . . Hao Yuan 、 Jiliang Tang 、 Xia Hu 和 Shuiwang Ji。。XGNN：面向图神经网络的模型级解释。载于*KDD '20*：第*26*届 *ACM SIGKDD* 知识发现与数据挖掘会议，线上活动，美国加利福尼亚州，*2020*年*8*月*23*日至*27*日，Rajesh Gupta、Yan Liu、Jiliang Tang 和 B. Aditya Prakash（编）。ACM，美国，430 438 10 1145 3394486 3403085 – 。https://doi.org/ . / . [61] Yuheng Zhang、Ruoxi Jia、Hengzhi Pei、Wenxiao Wang、Bo Li 和 Dawn Song。2020。秘密揭示者：针对深度神经网络的生成模型反演攻击。载于*2020*年 *IEEE/CVF* 计算机视觉与模式识别会议，*CVPR 2020*，美国华盛顿州西雅图，*2020*年*6*月*13*日至*19*日。计算机视觉基金会 / IEEE，美国，250 258 10 1109 42600 2020 00033 – 。[62] https://doi.org/ . /CVPR . . Xuejun Zhao、Wencan Zhang、Xiaokui Xiao 和 Brian Y. Lim。 。利用解释进行模型反演攻击。载于*2021*年 *IEEE/CVF* 国际计算机视觉会议，*ICCV 2021*，*17*, *2021* *10*662 672 10 1109 48922 2021加拿大魁北克省蒙特利尔，*10*月 -。IEEE，美国，– 。https://doi.org/ . /ICCV . . 00072 [63] Bolei Zhou、Aditya Khosla、Àgata Lapedriza、Aude Oliva 和 Antonio Torralba。2016。用于判别式定位的深层特征学习。载于*2016*年 *IEEE* 计算机视觉与模式识别会议，*CVPR 2016*，美国内华达州拉斯维加斯，*2016*年*6*月*27*日至*30*日。IEEE 计算机学会，美国，2921–2929。https: //doi.org/ 10 1109 2016 319 . /CVPR. . Luisa M. Zintgraf、Taco S. Cohen、Tameem Adel 和 Max Welling。2017。可视化深度神经网络决策：预测差异分析。载于第*5*届国际学习表征会议，*ICLR 2017*，法国土伦，*2017*年*4*月*24*日至*26*日，会议专题论文集。OpenReview.net，美国。https://openreview.net/forum?id=BJ5UeU9xx

机器学习即服务平台中的可解释人工智能

Azure 机器学习（Responsible AI） [2], Amazon SageMaker（Clarify） [47], Watson 机器学习（OpenScale） [11] 和 Google Cloud（Vertex 可解释人工智能） [10] 等主要机器学习即服务平台提供各种解释技术。这些平台提供的技术包括但不限于 SHAP [33], LIME [41], 置换特征重要性（PFI） [15],部分依赖图（PDP） [17],集成梯度 [54]和 XRAI [23]。然而，机器学习即服务提供商[2, 10, 47] 提供的大多数解释器主要是白盒的，主要面向模型开发者。它们能深入揭示模型的内部运作，与通常提供给查询用户的简单黑盒接口形成对比。这些应用场景包括但不限于偏差检测、公平性评估、异常值检测和降维。因此，对于商业化黑盒模型，解释器应针对给定样本提供效用，但不应揭示模型的全部内部运作。因此，所采用的解释器必须是局部的，最好与模型无关且属于事后解释。对于我们所考虑的模型，符合这些标准的解释器种类有限，包括 SHAP、LIME、反事实/假设情景（将在第8节讨论）以及特征重要性（LIME 和 SHAP 已提供）。