

基于对抗学习的强 PUF 安全结构研究

李艳, 刘威, 孙远路

(信息工程大学, 河南 郑州 450001)

摘要: 针对强物理不可复制函数 (PUF, physical unclonable function) 面临的机器学习建模威胁, 基于对抗学习理论建立了强 PUF 的对抗机器学习模型, 在模型框架下, 通过对梯度下降算法训练过程的分析, 明确了延迟向量权重与模型预测准确率之间的潜在联系, 设计了一种基于延迟向量权重的对抗样本生成策略。该策略与传统的组合策略相比, 将逻辑回归等算法的预测准确率降低了 5.4%~9.5%, 低至 51.4%。结合资源占用量要求, 设计了新策略对应的电路结构, 并利用对称设计和复杂策略等方法对其进行安全加固, 形成了 ALPUF (adversarial learning PUF) 安全结构。ALPUF 不仅将机器学习建模的预测准确率降低至随机预测水平, 而且能够抵御混合攻击和暴力破解。与其他 PUF 结构的对比表明, ALPUF 在资源占用量和安全性上均具有明显优势。

关键词: 物理不可复制函数; 对抗样本; 延迟向量; 对抗学习 PUF

中图分类号: TP393

文献标识码: A

doi: 10.11959/j.issn.2096-109x.2021019

Research on security architecture of strong PUF by adversarial learning

LI Yan, LIU Wei, SUN Yuanlu

Information Engineering University, Zhengzhou 450001, China

Abstract: To overcome the vulnerability of strong physical unclonable function, the adversarial learning model of strong PUF was presented based on the adversarial learning theory, then the training process of gradient descent algorithm was analyzed under the framework of the model, the potential relationship between the delay vector weight and the prediction accuracy was clarified, and an adversarial sample generation strategy was designed based on the delay vector weight. Compared with traditional strategies, the prediction accuracy of logistic regression under new strategy was reduced by 5.4% ~ 9.5%, down to 51.4%. The physical structure with low overhead was designed corresponding to the new strategy, which then strengthened by symmetrical design and complex strategy to form a new PUF architecture called ALPUF. ALPUF not only decrease the prediction accuracy of machine learning to the level

收稿日期: 2020-02-04; 修回日期: 2020-07-02

通信作者: 刘威, shivaree@163.com

基金项目: 国家自然科学基金 (61871405, 61802431)

Foundation Items: The National Natural Science Foundation of China (61871405, 61802431)

论文引用格式: 李艳, 刘威, 孙远路. 基于对抗学习的强 PUF 安全结构研究[J]. 网络与信息安全学报, 2021, 7(3): 115-122.
LI Y, LIU W, SUN Y L. Research on security architecture of strong PUF by adversarial learning [J]. Chinese Journal of Network and Information Security, 2021, 7(3): 115-122.

of random prediction, but also resist hybrid attack and brute force attack. Compared with other PUF security structures, ALPUF has advantages in overhead and security.

Keywords: strong physical unclonable function, adversarial sample, delay vector, adversarial learning PUF

1 引言

物理不可复制函数 (PUF, physical unclonable function) 作为一类独特的轻量级硬件安全机制, 以其防篡改、低资源占用等特点, 在密钥生成、身份识别和轻量级认证等领域拥有广阔的应用前景。

尽管 PUF 被定位为无法复制, 但研究表明 PUF 能够通过机器学习的方法建模。逻辑回归 (LR, logistic regression)、支持向量机 (SVM, support vector machine)、进化策略 (ES, evolution strategy)、深度学习 (DL, deep learning) 和集成学习 (EL, ensemble learning) 均能攻击一种或数种 PUF^[1-3]。为了应对机器学习建模带来的安全威胁, 研究者设计了多种 PUF 结构, 如异或仲裁器 PUF、混淆类 PUF 等, 但其中绝大多数被证明无法抵御机器学习攻击^[4-5]。机器学习已成为强 PUF 最大的安全威胁之一。

自 2005 年 Lowd 等首次提出对抗学习 (adversarial learning) 的概念以来, 机器学习算法自身的安全性就引起研究者的关注, 对抗机器学习逐渐成为人工智能领域的研究热点之一。目前, 针对深度神经网络的对抗攻击已经对人工智能在图像识别、自动驾驶等领域的应用构成了严重威胁^[6]。

为了抵御 PUF 的机器学习建模攻击, 本文基于对抗机器学习理论定义了强 PUF 的对抗机器学习模型, 在模型框架下设计了一种基于延迟向量权重的对抗策略。根据新策略设计了对抗学习 PUF (ALPUF, adversarial learning PUF) 电路结构, 并通过实验验证了新策略和 ALPUF 在抵御机器学习攻击时的优势。

2 PUF 建模与对抗机器学习

2.1 PUF 建模

仲裁器 PUF 是广泛使用的强 PUF 之一, 异或仲裁器 PUF、受控 PUF 和混淆类 PUF 等安全结构使用仲裁器 PUF 作为基本组件。

仲裁器 PUF 采用多路选择器级联结构, 如图 1

所示, 激励向量 C 的每一位对应控制一级多路器, 用于选择平行路径或者交叉路径, 输入信号沿两条不同路径传输, 最终由仲裁器比较传输时延, 形成响应 r 。激励 C 和响应 r 作为 PUF 的输入输出, 被合称为激励响应对 (CRP, challenge response pair)。

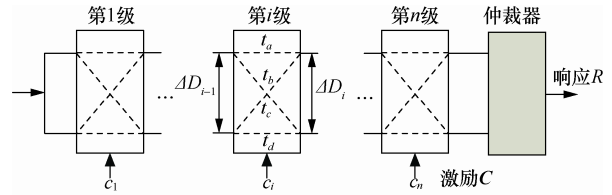


图 1 仲裁器 PUF 电路结构
Figure 1 PUF circuit structure of arbiter

设 $\delta_{c_i,i}$ 为输入为 c_i 时第 i 级时的延迟差, 则两路信号通过第 i 级后的总延迟差如式(1)所示。

$$\Delta D_i = \Delta D_{i-1}(1 - 2c_i) + \delta_{c_i,i} \quad (1)$$

响应 r 由最终的延迟差 ΔD_n 决定。

$$r = \begin{cases} 1, & \Delta D_n > 0 \\ 0, & \Delta D_n < 0 \end{cases} \quad (2)$$

分析表明仲裁器 PUF 可由 $2n$ 个延迟差值 δ 来描述, 经式(3)运算简化后只需 $n+1$ 个参数, 记为延迟向量 $W = (w_1, w_2, \dots, w_{n+1})$ 。

$$w_i = \begin{cases} \delta_{0,1} - \delta_{1,1}, & i = 1 \\ \delta_{0,i-1} + \delta_{1,i-1} + \delta_{0,i} - \delta_{1,i}, & 1 \leq i \leq n-1 \\ \delta_{0,n} + \delta_{1,n}, & i = n \end{cases} \quad (3)$$

延迟差 ΔD_n 可以由延迟向量 W 与特征向量 Φ 的标量乘积得到, 计算公式为 $\Delta D_n = W^T \Phi$, 特征向量由激励向量 C 计算得到, 如式(4)所示。

$$\phi_i = \begin{cases} \prod_{l=i}^n (1 - 2c_l), & 1 \leq i \leq n \\ 1, & i = n+1 \end{cases} \quad (4)$$

延迟向量 W 可唯一描述仲裁器 PUF。延迟向量 W 的分量 $w_1, w_2, \dots, w_{n+1}$ 由 PUF 制造过程决定, 取值各不相同, 因此本文将 $|w_i|$ 定义为各延迟分量的权重。虽然 PUF 的延迟参数无法通过测量直接得到, 但是获取足够数量的 CRP 后, 可用机器

学习方法训练得出延迟向量, 建立数学模型。

2.2 对抗机器学习

2006年, 机器学习系统的攻击分类和对抗建模的概念首次明确被提出, Biggio 等^[7]在 2014 年进一步完善了机器学习的对抗模型, 提出从对抗目标、对抗知识、对抗能力、对抗策略 4 方面来建立模型。早期的对抗方法主要针对 SVM、朴素贝叶斯、聚类 and 特征选择等机器学习方法^[8], 而目前研究热点主要针对人工神经网络。

按照对机器学习模型造成的安全损害可以将对抗机器学习方法分为完整性攻击、可用性攻击和隐私窃取攻击; 按照攻击目标阶段可分为针对训练过程的攻击和针对测试/推理过程的攻击。投毒攻击属于针对训练过程的攻击, 攻击者通过注入一些精心伪造的对抗样本(adversarial sample), 破坏原有的训练数据的概率分布, 从而使训练出的模型的分类或者聚类精度降低, 达到破坏训练模型的目的。

对抗样本指的是一类人为构造的样本, 通过对原始样本添加特定的扰动, 使分类模型对新构造的样本产生错误的判断。对抗样本生成算法主要有快速梯度(fast gradient sign method)攻击和雅克比映射攻击(Jacobian-based saliency map attack)和深度欺骗(deep fool)攻击^[9]。快速梯度攻击的主要思想是寻找机器学习模型的梯度变化最大的方向, 按照此方向添加扰动, 导致模型进行错误的分类。

2.3 抗机器学习的 PUF 研究现状

从早期的异或 PUF、受控 PUF 到目前广泛研究的混淆类 PUF, 研究人员为对抗机器学习建模设计了多种 PUF 结构。但是, 异或 PUF 和受控 PUF 被证明易受到机器学习和侧信道混合攻击^[10]。2019 年, Delvaux^[5]对新提出的 PUF 安全结构进行了分析, 针对性地设计了攻击方法, 指出混淆类 PUF 难以抵御机器学习建模攻击。

对抗机器学习在 PUF 安全领域的应用研究相对较少, 2019 年, Wang 等^[11]提出了针对 PUF 建模的对抗攻击方法, 设计了 PUF 投毒攻击的机制, 实验结果表明, 基于投毒的对抗攻击将 LR 的预测准确率由 99.89%降至 56.7%, 将 CMA-ES 的预测准确率由 99.97%降至 65.68%, 但对抗样

本生成策略主要基于设计者的经验, 缺乏相关的理论分析, 且实验测试对象规模相对较小, 只涵盖了 64 bit 的 PUF。

2018 年, Prado 等^[12]将 PUF 的所有 CRP 按汉明重量分类后按类别对外提供, 通过 CRP 的对抗性选择将 AdaBoost 算法的预测准确率由 89%降至 74%, 但并未提及该策略的物理实现方法。

利用对抗学习提高 PUF 的建模难度, 降低机器学习模型的预测准确率, 是保护 PUF 应用安全的有效途径。但目前的研究尚存在对抗样本生成策略缺乏理论支持、对抗效果未达到最优性能、物理实现方式考虑不足等问题。

3 基于延迟向量权重的对抗样本生成

3.1 强 PUF 的对抗机器学习模型

在对抗学习领域, 针对分类问题的对抗样本生成策略通常基于梯度上升算法, 通过运算不断调整样本中的输入值, 最终生成最优的有毒样本。但是, 目前的对抗模型和样本生成算法无法适用于生成 PUF 对抗样本, 主要原因有: PUF 的激励 C 是布尔向量, 无法在其原有值上再叠加较小的数值, 只能做翻转操作; 攻击模型中通常设定攻击者可以物理接触 PUF, 激励 C 由攻击者选定, 样本中的激励无法修改。因此, 需要重新建立适用于 PUF 安全领域的对抗模型, 在模型框架下设计新样本生成策略。

基于 Biggio 的对抗机器学习理论, 结合强 PUF 的特性、应用场景和攻击模型设定, 给出强 PUF 对抗机器学习模型。

模型 1 强 PUF 对抗机器学习模型

对抗目标: 以破坏机器学习算法的可用性为主要目的, 通过向 PUF 的 CRP 集中加入伪造的 CRP, 即有毒数据, 在训练阶段阻碍机器学习算法生成正确训练模型。

具备知识: 对抗者作为 PUF 安全防护方, 知悉 PUF 相关的所有知识, 包括 PUF 模型、训练数据、PUF 延迟向量等。

拥有能力: 对抗者可以对 PUF 结构、物理实现等进行改进, 具备在机器学习算法运行前修改训练集、投放有毒数据的能力, 且不受传统对抗方法对训练集中有毒数据所占比例的限制(传统

方法为了达到最优性能,有毒数据在训练数据集中比例通常小于 20%)。

对抗策略:以大幅降低机器学习建模方法的预测准确率为目的,兼顾实现时的资源占用量设计最优策略,根据激励各位的组合运算结果调整响应值,生成有毒 CRP。PUF 的有毒数据生成与传统对抗方式不同,在攻击场景中,机器学习建模攻击者被认为能物理接触 PUF 并拥有控制权,激励由攻击者选定,对抗者无法对激励进行调整,因此只能通过修改响应的方式篡改真实 CRP,达到降低模型预测准确率的目的。

强 PUF 对抗机器学习模型要求使用者掌握目标 PUF 的结构、参数信息和训练数据,通过这些信息构造对抗样本,属于白盒模型。该模型通过产生有毒 CRP 的方式,增加机器学习算法预测 PUF 参数的难度,降低算法预测准确率,具有一定的普适性。

3.2 对抗样本生成

生成合适的对抗样本是决定对抗效果的关键。PUF 建模属于机器学习中的分类问题,逻辑回归、支持向量机和深度学习等方法被广泛应用于仲裁器类 PUF 的建模。其中,LR 是众多 PUF 攻击方法中研究最多、使用最广泛的方法之一。逻辑回归攻击仲裁器 PUF,按照式(6)对训练集 M 中的 CRP 使用梯度下降法,学习出延迟向量即完成建模。

$$\hat{W} = \arg \min_{W} l_{(C,r) \in M}(M, W) \quad (6)$$

逻辑回归的本质是用线性回归的预测结果逼近对数概率 $\ln \frac{r}{1-r}$ 。用梯度下降法更新参数如式(7)所示。

$$W \leftarrow W + \eta \left(r_i - \frac{e^{\vec{w}^T \vec{\phi}_i}}{1 + e^{\vec{w}^T \vec{\phi}_i}} \right) \vec{\phi}_i \quad (7)$$

特征向量 Φ 由激励向量 C 运算得到,根据 PUF 对抗学习模型,对抗者无法修改激励向量 C ,也就无法改变 Φ ,对抗样本只能以修改响应 r 的方式生成。翻转响应 r 将使参数 W 向错误方向调整更新,根据式(7),更新对 W 中各项产生相同影响,使参数向错误方向更新相同值。

响应 r 的翻转条件是策略设计的主要内

容,翻转条件可以与激励 C 无关,也可以是 C 中一位或多位的运算结果,设计时主要遵循以下原则。

原则 1 选择权重小的延迟分量所对应激励位作为基本元素,可以获得更好的对抗效果。原因是尽管 W 中各延迟分量同步更新的数值完全相同,但对 W 中绝对值较小的分量来说,该数值产生的影响更大,而对绝对值较大的分量,错误更新的影响则相对较小。

原则 2 策略中同一激励位触发翻转的条件需保持一致。即如设定某激励位 c_i 为 1 时触发响应翻转,则不能同时将该位为 0 作为触发翻转条件。原因是同时作为翻转条件意味着所有 CRP 均翻转,这将导致机器学习算法训练得到只是参数 W 符号相反的正确模型。

以针对 LR 和 SVM 生成对抗数据为例,分别选取 W 中绝对值最小、最大和中位分量所对应的激励位,当对应的 ϕ_i 取值为 1 时翻转响应 r ,训练样本集包含 100×10^3 对 CRP,其中对抗样本占 50%,对抗机器学习的效果如表 1 所示。

表 1 延迟向量权重与策略对抗效果的关系
Table 1 Relationship between delay vector weight and strategy countermeasure effect

W 取值	预测准确率	
	LR	SVM
-0.021 266 21 (Min)	51.4%	51.7%
2.323 293 36 (Max)	62.3%	62.5%
1.042 470 09 (Avg)	56.65%	56.9%
Nul	99.8%	99%

使用对抗样本后,LR 的预测准确率由未使用时的 99.8%下降至 62.3%以下,SVM 的预测准确率由未使用时的 99%下降至 62.5%以下。其中,选择 W 中权重小的分量所对应激励位作为翻转条件,可以进一步将预测准确率降至 51.4%和 51.7%。

基于上述分析,对 128 bit 仲裁器 PUF 按照式 $r' = \text{PUF}(C) \oplus \phi_i$ 生成对抗样本,依次从第 0 位到第 128 位生成训练集 (Φ 比 C 多一位),提供给 LR 进行建模,得到预测准确率曲线,在图 2 中用橙色曲线表示。图中蓝色曲线为 $|W|$,即 W 的绝对值,为了方便对比,对 $|W|$ 进行了归一化处理。

从图2中可以看出, 预测准确率同 $|W|$ 呈严格正相关。因此, 延迟向量权重可以作为响应翻转条件, 用于对抗样本生成策略设计。

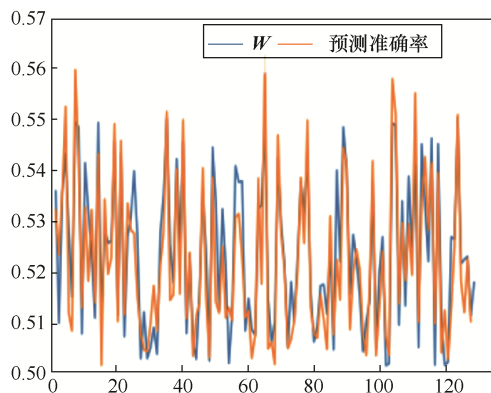


图2 W 权重与预测准确率对比

Figure 2 Comparison of W weight and prediction accuracy

算法 1 基于延迟向量权重的对抗样本生成算法

输入 原始激励响应对集合 $\{(C, r)\}$

输出 对抗样本集合 $\{(C, r')\}$

1) 特征向量 $\Phi \leftarrow$ 激励向量 C

2) $\hat{W} \leftarrow \arg \min_{W} LR_{(\Phi, r) \in M}(M)$; 使用 LR 等算

法针对训练集 $\{(\Phi, r)\}$ 进行学习, 得到延迟向量 W

3) $W_{\text{weight}} \leftarrow \text{sort}(|W|)$; 对 W 中的各分量按绝对值从小到大排序

4) $i, j \dots \leftarrow W_{\text{weight}}$ 的分量序号

5) $r' = \text{PUF}(C) \oplus \phi_i$ 或 $r' = \text{PUF}(C) \oplus f(\phi_i, \phi_j \dots)$

(其中 $f()$ 为布尔函数), 生成对抗样本集合 $\{(C, r')\}$

3.3 生成策略性能对比

文献[11]中提出了随机翻转、组合翻转等对抗样本生成策略, 这些策略与本文策略的最大区别是没有参考延迟向量 W 的权重。

随机翻转策略是将响应按照一定比例随机翻转, 翻转行为和 PUF 的激励无关, 由设置的随机数发生器决定。

$$r' = \text{PUF}(C) \oplus \text{rand}$$

组合翻转策略是将激励的不同位进行逻辑运算, 由运算得到的最终结果决定是否翻转响应。

$$r' = \text{PUF}(C) \oplus f(\phi_i, \phi_j \dots)$$

测试中选用 $f = \phi_i \oplus \phi_j \oplus \phi_m \oplus \phi_n$, 组合策略

1 和策略 2 的区别在于随机选择对应的序号和选择权重较大分量对应的序号。

在相同条件下对 4 种策略进行测试, 结果如表 2 所示, 可以看出, 随机翻转性能最好, 其次是基于延迟向量权重的生成策略性能, 基本可以使 LR 和 SVM 的预测准确率降低至随机预测的水平, 组合翻转策略的性能随着参与逻辑运算位所对应的延迟向量权值变化。

表 2 不同对抗样本生成策略的效果对比

Table 2 Comparison of the effects of different counter sample generation strategies

策略	预测准确率	
	LR	SVM
本文	51.4%	51.7%
随机翻转 ^[11]	50.2%	49.7%
组合翻转 1 ^[11]	56.8%	57.4%
组合翻转 2 ^[11]	60.9%	61.3%

随机策略尽管性能最好, 但其抵御机器学习的能力来自随机数发生器, 对手无法攻破的同时, 使用者也难以判定, 因此难以实用。组合策略中选取了不同权重延迟参数对应的激励位, 根据原则 1 判断, 其性能介于最大权重和最小权重对应预测准确率之间。

根据对抗样本的可转移性理论, 被一种机器学习算法错分类的对抗样本对其他算法也适用, 使用本策略生成对抗样本, 也能够降低其他机器学习算法的预测准确率。表 2 的数据结果显示, 4 种策略对 SVM 和 LR 两种算法都显示出了性能近似的对抗效果。

3.4 投毒比例影响分析

按照强 PUF 对抗学习模型, 对抗者可向训练集中投放任意比例的有毒数据, 根据 PAC 理论, 有毒数据比例对机器学习算法的预测准确率具有决定性的影响。

以 LR 攻击仲裁器 PUF 为例, 分别使用新策略和随机翻转策略在训练集中加入 12.5%、25% 和 50% 的有毒数据, 训练集规模为 100×10^3 , 机器学习算法的预测准确率如表 3 所示。

根据实验结果可以看出, 预测准确率受到投毒比例和对抗数据生成策略两方面的影响, 投毒比例很大程度上决定了分类算法的预测准确率。

但在相同比例的情况下,不同策略的对抗效果存在较大差距。例如,分别采用组合翻转和本文策略生成有毒数据占 25% 的训练集对 LR 算法进行训练,得到模型的预测准确率分别是 78.2% 和 73%。文献[11]中所论述的不同组合策略对预测准确率的影响,其实主要由投毒比例不同所导致。

表 3 训练集成分和对抗策略对预测准确率的影响
Table3 Effects of training set composition and countermeasures on prediction accuracy

投毒比例	预测准确率	
	基于延迟向量权重	组合翻转
0%	99.8%	99.8%
12.5%	86.02%	90.6%
25%	73%	78.2%
50%	51.4%	56.8%

4 对抗学习 PUF

4.1 基于对抗策略的 PUF 结构

对抗样本生成策略必须能在满足低资源占用量的前提下形成物理电路结构,才能够保证策略能够推广应用。

特征向量 Φ 由激励向量 C 计算得到。如果严格按照对抗策略设计 PUF 结构,根据式(5),实现 $r' = \text{PUF}(C) \oplus \phi_i$ 需要多个激励位的连续 $n-i$ 次异或,硬件开销过大。

由式(5)可知, Φ 中的 $\phi_{i+1} \sim \phi_{n+1}$ 和 c_i 无关, c_i 的影响范围是 $\phi_1 \sim \phi_i$ 。其中, c_i 直接影响 ϕ_i 的值,即 c_i 改变时 ϕ_i 一定发生翻转, c_i 对 $\phi_{i-1} \sim \phi_1$ 的影响依次递减。例如, ϕ_{i-1} 是否翻转由 c_i 和 c_{i-1} 共同决定。考虑到策略实现时的资源占用量,将 ϕ_i 调整为关系最为密切的 c_i ,则对应策略修改为 $r' = \text{PUF}(C) \oplus c_i$,激励中的 1 位直接与响应做异或操作,可将硬件开销降至最低,实现的电路结构如图 3 所示。

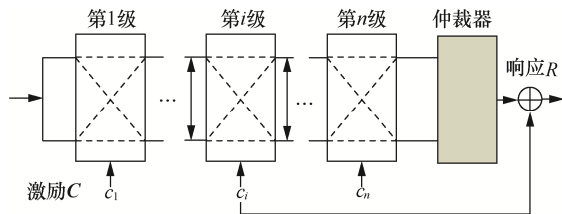


图 3 基于对抗策略的 PUF 电路结构
Figure 3 PUF circuit structure based on countermeasure strategy

基于 Xilinx ZC702 开发板实现该电路结构,依次改变激励向量 C 中的每一位测试得到训练集,LR 学习建模的预测准确率曲线如图 4 所示,可以看出,将策略中 ϕ_i 调整为 c_i 后,预测准确率也跟随 $|W|$ 的权重变化,基于特征向量权重的对抗数据生成策略仍然有效。

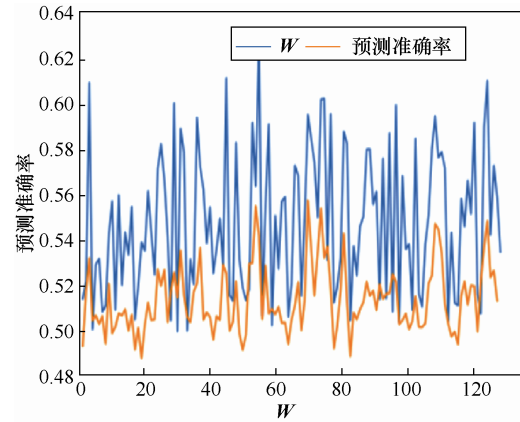


图 4 按 C 测试 W 权重与预测准确率对比
Figure 4 Test W weight and prediction accuracy according to C

4.2 抗复杂攻击的对抗学习 PUF

机器学习并非强 PUF 的唯一威胁,一些复杂的攻击方法仍可能威胁到基于对抗策略的 PUF。例如,多数抗机器学习建模的 PUF 结构被证明难以抵御的机器学习/侧信道混合攻击^[13]。以异或 PUF 为例,因为构成异或 PUF 的仲裁器 PUF 用触发器实现仲裁,响应为“1”时触发器发生翻转,其功耗远高于响应为“0”不发生翻转时。测量供电引脚上的电压/电流值,就能够确定响应值。测量得到响应为“1”的仲裁器总数后,可以使用机器学习对异或 PUF 建模。

混合攻击也能成功攻击对抗 PUF。在功耗分析面前,基于对抗策略的 PUF 的真实响应值无法隐藏,攻击者可以轻易剔除或者转化有毒数据。为了防止混合攻击,可以通过增加电路的对称性以防止侧信道信息泄露。在实现时对应设置两个触发器,两条路径的信号分别进入两个触发器,形成两个相反的值,无论响应为何值,功耗都一样,以此防御混合攻击。

图 3 中的 PUF 结构无法抵御暴力攻击。尽管攻击者不知道 c_i 的具体序号,但可以穷举测试激励 C 的每一位,最终找到正确模型。对抗暴力破

解的方法是目的是让 PUF 结构更复杂,可能出现的情况数量更多,运算量超过攻击者的计算能力。因此,PUF 安全结构中不能局限于使用单个激励位,而应该按照延迟向量权重顺序同时选定更多的激励位参与布尔函数的运算,布尔函数也不局限于固定算法。按照上述两项改进,基于对抗策略设计的安全 PUF 为 ALPUF,其电路结构如图 5 所示。ALPUF 不仅能够抵御单独的机器学习建模,而且能够抗机器学习/侧信道混合攻击和暴力破解。

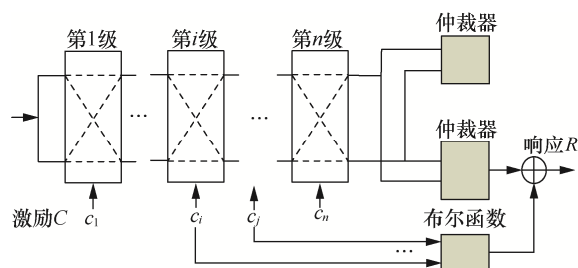


图5 ALPUF 电路结构
Figure 5 ALPUF circuit structure

与近年来提出的 PUF 安全结构对比,ALPUF 在对抗性能和资源占用量等方面均具有优势。表 4 中列出了 PUF 安全性和硬件资源占用上的对比。资源占用以基本的仲裁器 PUF 为标准,列出实现对应结构额外所需的组件,从表 4 中可以看出,ALPUF 在仲裁器 PUF 的基础上只增加了 3 个异或逻辑,资源占用量最低。安全性评估中使用逻辑回归算法攻击所列出的 PUF,对比预测准确率。其中,对 ALPUF 的预测准确率仅为 51.4%,接近随机预测,远低于其他 PUF 结构,而且 ALPUF 的对称结构能够抵御功率混合攻击,由此可以看出 ALPUF 具备更强的对抗机器学习攻击的能力。

ALPUF 不仅能单独使用,而且可以作为基本

组件,替代仲裁器 PUF 参与组成复杂的 PUF 安全结构。例如,将异或 PUF 中的仲裁器 PUF 换成 ALPUF,可进一步提高其抗机器学习的能力。以 2-XOR PUF 为例,相同条件下使用 LR 建模的预测准确率为 99.2%,将组件替换为 ALPUF 后,LR 建模的预测准确率下降至随机预测水平。

5 结束语

为了提高强 PUF 抵御机器学习攻击的能力,本文基于对抗学习理论定义了强 PUF 对抗学习模型。在模型框架下,本文设计了基于延迟向量权重的对抗样本生成策略,按照新策略设计了 ALPUF 电路结构。实验结果表明,与现有策略相比,使用新策略生成的对抗样本能将 LR、SVM 等分类算法的预测准确率降至随机预测水平;ALPUF 在低资源占用的前提下,能够抵御复杂的机器学习建模攻击。

参考文献:

- [1] SAHOO D P, NGUYEN P H, MUKHOPADHYAY D, et al. A case of lightweight PUF constructions: cryptanalysis and machine learning attacks[J]. IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, 2015, 34(8): 1334-1343.
- [2] LIU Y, XIE Y, BAO C, et al. A combined optimization-theoretic and side-channel approach for attacking strong physical unclonable functions[J]. IEEE Transactions on Very Large Scale Integration Systems, 2017, 26(1): 73-81.
- [3] RÜHRMAIR U, SÖLTER J, SEHNKE F, et al. PUF modeling attacks on simulated and silicon data[J]. IEEE Transactions on Information Forensics & Security, 2013, 8(11): 1876-1891.
- [4] YE J, GUO Q, HU Y, et al. Modeling attacks on strong physical unclonable functions strengthened by random number and weak PUF[C]//2018 IEEE 36th VLSI Test Symposium. 2018: 1-6.
- [5] DELVAUX J. Machine-learning attacks on PolyPUFs, OB-PUFs, RPUFs, LHS-PUFs, and PUF-FSMs[J]. IEEE Transactions on Information Forensics and Security, 2019, 14(8): 2043-2058.
- [6] 宋蕾, 马春光, 段广晗. 机器学习安全及隐私保护研究进展[J].

表 4 PUF 安全性及硬件资源占用上的对比
Table 4 Comparison of PUF security and resource utilization

类型	预测准确率 (基于 LR)	对抗功率混合攻击	资源占用量
AL PUF	51.4%	√	3 XOR
4-XOR PUF	98.2%	×	3 XOR 3 APUF
OB PUF ^[14]	72%	√	1 TRNG3 APUF
R PUF ^[15]	80%	√	1 TRNG 1 LFSR
控制型 PUF ^[16]	69.8%	√	1 XOR 1 TRNG 1 混淆模块

网络与信息安全学报, 2018, 4(8): 1-11.

SONG L, MA C G, DUAN G H. Machine learning security and privacy: a survey[J]. Chinese Journal of Network and Information Security, 2018, 4(8): 1-11.

- [7] BIGGIO B, NELSON B, LASKOV P. Poisoning attacks against support vector machines[C]//ICML'12.2012. 1467-1474.
- [8] JAGIELSKI M, OPREA A, BIGGIO B, et al. Manipulating machine learning: poisoning attacks and countermeasures for regression learning[C]//2018 IEEE Symposium on Security and Privacy. 2018. 19-35.
- [9] BIGGIO B, ROLI F. Wild patterns: ten years after the rise of adversarial machine learning[J]. Pattern Recognition, 2018, 84:317-331.
- [10] BECKER G T. The gap between promise and reality: on the insecurity of XOR arbiter PUFs[C]//Cryptographic Hardware and Embedded Systems. 2015: 535-555.
- [11] WANG S J, CHEN Y S, KATHERINE L, et al. Adversarial attack against modeling attack on PUFs[C]//DAC2019. 2019:1-6.
- [12] PRADO H, PATIL C. Defeating strong PUF modeling attack via adverse selection of challenge-response pairs[C]//AsianHOST2018. 2018: 25-30.
- [13] XU X, MAHMOUD A, MAJZOBI M, et al. Efficient power and timing side channels for physical unclonable functions[C]//International Workshop on Cryptographic Hardware and Embedded Systems. 2014: 476-492.
- [14] GAO Y, LI G, MA H, et al. Obfuscated challenge-response: a secure lightweight authentication mechanism for PUF-based pervasive devices[C]//Workshop in Conjunction with IEEE International Conference on Pervasive Computing and Communications. 2016: 1-6.
- [15] JING Y, YU H, LI X. RPUF: physical unclonable function with randomized challenge to resist modeling attack[C]//Hardware-oriented Security & Trust. 2016: 1-6.
- [16] 徐金甫, 吴缙, 李军伟. 基于敏感度混淆机制的控制型物理不可

复制函数研究[J]. 电子与信息学报, 2019, 41(7): 1601-1609.

XU J F, WU J, LI J W. Research on the control physical non-clonable function based on sensitivity cofounding mechanism[J]. Journal of Electronics and Information Technology, 2019, 41(7): 1601-1609.

[作者简介]



李艳 (1988-) 女, 河南郑州人, 信息工程大学讲师, 主要研究方向为网络空间安全。



刘威 (1982-) 男, 湖北随州人, 信息工程大学讲师, 主要研究方向为硬件安全。



孙远路 (1974-) 男, 河南商丘人, 信息工程大学副教授, 主要研究方向为管理科学与工程。