(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 109951484 A
(43)申请公布日 2019.06.28

(21)申请号 201910213571.6

(22)申请日 2019.03.20

(71)申请人 四川长虹电器股份有限公司
地址 621000 四川省绵阳市高新区绵兴东路35号

(72)发明人 钟倩

(74)专利代理机构 四川省成都市天策商标专利事务所 51213
代理人 李洁

(51)Int.Cl.
*H04L 29/06*(2006.01)

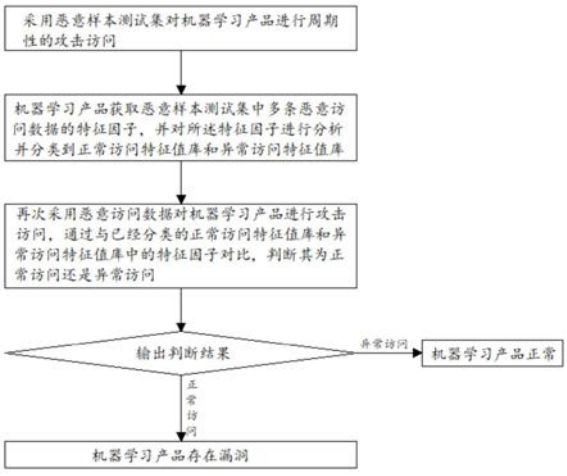
权利要求书1页 说明书3页 附图1页

(54)发明名称

针对机器学习产品进行攻击的测试方法及系统

(57)摘要

本发明公开了一种针对机器学习产品进行攻击的测试方法,包括以下步骤:采用恶意样本测试集对机器学习产品进行周期性的攻击访问;机器学习产品获取恶意样本测试集中多条恶意访问数据的特征因子,并对所述特征因子进行分析,根据分析结果对所述特征因子进行分类管理;再次采用恶意样本测试集中的一条恶意访问数据对机器学习产品进行攻击访问,机器学习产品对该条恶意访问数据中的特征因子进行分析,并与已经分类的正常访问特征值库和异常访问特征值库中的特征因子进行对比,判断其为正常访问还是异常访问;输出判断结果;本发明还公开了一种针对机器学习产品进行攻击的测试系统,本发明进一步提高了机器学习产品的安全性。




CN 109951484 A

1.一种针对机器学习产品进行攻击的测试方法,其特征在于,包括以下步骤:

步骤一、采用恶意样本测试集对机器学习产品进行周期性的攻击访问,所述恶意测试样本集中包括多条恶意访问数据;

步骤二、机器学习产品获取恶意样本测试集中多条恶意访问数据的特征因子,并对所述特征因子进行分析,根据分析结果对所述特征因子进行分类管理,判断将其放入正常访问特征值库还是异常访问特征值库;

步骤三、再次采用恶意样本测试集中的一条恶意访问数据对机器学习产品进行攻击访问,机器学习产品对该条恶意访问数据中的特征因子进行分析,并与已经分类的正常访问特征值库和异常访问特征值库中的特征因子进行对比,判断其为正常访问还是异常访问;

步骤四、输出判断结果。

2.根据权利要求1所述的针对机器学习产品进行攻击的测试方法,其特征在于,所述步骤四中,若输出的判断结果为正常访问,则说明该机器学习产品存在漏洞,若输出的判断结果为异常访问,则说明该机器学习产品正常。

3.根据权利要求1所述的针对机器学习产品进行攻击的测试方法,其特征在于,所述步骤一中,对机器学习产品进行攻击访问的恶意访问数据的数量大于同一时期机器学习产品的正常访问量。

4.根据权利要求1所述的针对机器学习产品进行攻击的测试方法,其特征在于,所述步骤一中,恶意样本测试集对机器学习产品攻击访问的周期大于该机器学习产品防御算法的周期。

5.一种针对机器学习产品进行攻击的测试系统,其特征在于,包括:

样本模块,用于存储进行攻击访问的恶意样本测试集并训练恶意样本集中恶意访问数据的特征因子;

算法模块,用于机器学习产品获取恶意访问数据的特征因子,并对获取的特征因子进行分析与分类,以及预测判断模块的建立与完善;

预测判断模块,包括正常访问特征值库和异常访问特征值库,用于存储分类后的特征因子,实时监控机器学习产品服务器上的访问日志,并与正常访问特征值库和异常访问特征值库已经存储的特征因子进行对比,发现异常访问行为则进行预测输出;

恶意访问模块,用于采用一条恶意访问数据对机器学习产品进行攻击访问,该恶意访问数据中包括待预测的特征因子;

预测结果模块,用于接收预测判断模块的预测输出结果,并判断恶意访问模块中的恶意访问数据对机器学习产品进行攻击访问的测试是否成功。

6.根据权利要求5所述的针对机器学习产品进行攻击的测试系统,其特征在于,所述的恶意样本集为人工提供的恶意访问数据。

# 针对机器学习产品进行攻击的测试方法及系统

## 技术领域

[0001] 本发明涉及互联网安全测试技术领域，特别是一种针对机器学习产品进行攻击的测试方法及系统。

## 背景技术

[0002] 随着互联网的快速发展，互联网的安全现如今已愈加被人们所重视，互联网的攻击和防护手段也在逐渐升级，机器学习的兴起与运用给安全测试带来了一大挑战。经过长期的发展，机器学习的运用已在各个领域崭露头角。由于机器学习所做的网络应用模块还未有效果显著的测试方法，一种网络攻击者利用机器学习产品进行攻击的测试方法就应运而生了。

[0003] 目前在进行机器学习产品的安全测试的时候，存在测试方法传统、单一、缺乏显著的测试结果等局限性。

## 发明内容

[0004] 为解决现有技术中存在的问题，本发明的目的是提供一种针对机器学习产品进行攻击的测试方法及系统，本发明通过搜集大量的恶意访问数据对使用机器学习的目标网络进行访问，促使恶意样本集中的恶意访问数据变成机器学习的判断规则并绕过机器学习的检测规则，导致机器学习系统被污染，使攻击者有机可乘；安全研究员预测机器学习最终将基于测试方法和结果来实时修改代码以避免此危害。

[0005] 为实现上述目的，本发明采用的技术方案是：一种针对机器学习产品进行攻击的测试方法，包括以下步骤：

[0006] 步骤一、采用恶意样本测试集对机器学习产品进行周期性的攻击访问，所述恶意测试样本集中包括多条恶意访问数据；

[0007] 步骤二、机器学习产品获取恶意样本测试集中多条恶意访问数据的特征因子，并对所述特征因子进行分析，根据分析结果对所述特征因子进行分类管理，判断将其放入正常访问特征值库还是异常访问特征值库；

[0008] 步骤三、再次采用恶意样本测试集中的一条恶意访问数据对机器学习产品进行攻击访问，机器学习产品对该条恶意访问数据中的特征因子进行分析，并与已经分类的正常访问特征值库和异常访问特征值库中的特征因子进行对比，判断其为正常访问还是异常访问；

[0009] 步骤四、输出判断结果。

[0010] 作为一种优选的实施方式，所述步骤四中，若输出的判断结果为正常访问，则说明该机器学习产品存在漏洞，若输出的判断结果为异常访问，则说明该机器学习产品正常。

[0011] 作为另一种优选的实施方式，所述步骤一中，对机器学习产品进行攻击访问的恶意访问数据的数量大于同一时期机器学习产品的正常访问量。

[0012] 作为另一种优选的实施方式，所述步骤一中，恶意样本测试集对机器学习产品攻

击访问的周期大于该机器学习产品防御算法的周期。

[0013] 本发明还提供一种针对机器学习产品进行攻击的测试系统,包括:

[0014] 样本模块,用于存储进行攻击访问的恶意样本测试集并训练恶意样本集中恶意访问数据的特征因子;

[0015] 算法模块,用于机器学习产品获取恶意访问数据的特征因子,并对获取的特征因子进行分析与分类,以及预测判断模块的建立与完善;

[0016] 预测判断模块,包括正常访问特征值库和异常访问特征值库,用于存储分类后的特征因子,实时监控机器学习产品服务器上的访问日志,并与正常访问特征值库和异常访问特征值库已经存储的特征因子进行对比,发现异常访问行为则进行预测输出;

[0017] 恶意访问模块,用于采用一条恶意访问数据对机器学习产品进行攻击访问,该恶意访问数据中包括待预测的特征因子;

[0018] 预测结果模块,用于接收预测判断模块的预测输出结果,并判断恶意访问模块中的恶意访问数据对机器学习产品进行攻击访问的测试是否成功。

[0019] 作为一种优选的实施方式,所述的恶意样本集为人工提供的恶意访问数据。

[0020] 本发明的有益效果是:本发明从网络攻击者利用机器学习产品进行攻击的测试,角度新颖,能从绕过机器学习算法这一角度攻击被测试的机器学习产品,给安全研究员提供了一种新的测试思路,为即时发现机器学习产品漏洞提供有效手段,安全研究员最终将基于测试方法和结果来实时修改代码以避免此危害,保证了产品和用户的信息安全,进一步提高了机器学习产品的安全性;通过本发明的测试,发现基于DQN(DeepQ-Learning)算法(深度学习deeplearning与强化学习reinforcementlearning相结合)、TRPO(Trust Region Policy Optimization)信赖域策略优化以及A3C(Actor-Critic Algorithm)算法的机器学产品均被恶意访问数据攻击成功,说明基于此类算法的机器学习产品还存在漏洞,方便安全研究员修改代码提高机器学习产品的安全性。

## 附图说明

[0021] 图1为本发明实施例的流程框图;

[0022] 图2为本发明实施例的系统框图。

## 具体实施方式

[0023] 下面结合附图对本发明的实施例进行详细说明。

[0024] 实施例:

[0025] 如图1所示,一种针对机器学习产品进行攻击的测试方法,包括以下步骤:

[0026] 预置条件:被攻击的机器学习产品服务器(以防火墙为例)、操作系统为WIN7或WIN8或WIN10或Linux的PC;

[0027] 步骤一、采用恶意样本测试集对机器学习产品进行周期性的攻击访问,所述恶意测试样本集中包括多条恶意访问数据;其中,对机器学习产品进行攻击访问的恶意访问数据的数量大于同一时期机器学习产品的正常访问量,恶意样本测试集对机器学习产品攻击访问的周期大于该机器学习产品防御算法的周期;

[0028] 步骤二、机器学习产品获取恶意样本测试集中多条恶意访问数据的特征因子,并

对所述特征因子进行分析，根据分析结果对所述特征因子进行分类管理，判断将其放入正常访问特征值库还是异常访问特征值库；在本实施例中，大量的恶意访问数据通过长时间访问被测试的机器学习产品，使得机器学习产品获取大量的恶意访问数据信息，并且分析所有的恶意访问数据，由于恶意访问数据的数量巨大，机器学习产品将恶意访问数据判断为正常访问数据，并提取它们的特征因子放入正常访问特征值库中；

[0029] 步骤三、再次采用恶意样本测试集中的一条恶意访问数据(例如:<script>alert('xss');</script>)对机器学习产品进行攻击访问，机器学习产品对该条恶意访问数据中的特征因子进行分析，并与已经分类的正常访问特征值库和异常访问特征值库中的特征因子进行对比，判断其为正常访问还是异常访问；

[0030] 步骤四、输出判断结果；本实施例中输出为正常访问数据，说明此恶意访问数据可以以正常数据的身份进行访问，成功的绕过了机器学习产品的WAF(Web Application Firewall)防火墙系统，并弹出XSS的弹框，攻击者攻击成功，通过该方法对机器学习产品的测试结果说明该机器学习产品存在漏洞。

[0031] 如图2所示，本实施例还提供一种针对机器学习产品进行攻击的测试系统，包括：

[0032] 样本模块，用于存储进行攻击访问的恶意样本测试集并训练恶意样本集中恶意访问数据的特征因子；

[0033] 算法模块，用于机器学习产品获取恶意访问数据的特征因子，并对获取的特征因子进行分析与分类，以及预测判断模块的建立与完善；本实施例中，算法模块中攻击者通过长时间，大量的访问被测试的机器学习产品，使得算法模块获取大量的恶意访问数据信息，并且分析所有的恶意访问数据，由于恶意访问数据的数量巨大，算法模块判断恶意数据为正常访问数据，并提取它们的特征值放入预测判断模块；

[0034] 预测判断模块，包括正常访问特征值库和异常访问特征值库，用于存储分类后的特征因子，实时监控机器学习产品服务器上的访问日志，并与正常访问特征值库和异常访问特征值库已经存储的特征因子进行对比，发现异常访问行为则进行预测输出；本实施例中，当再次采用恶意样本测试集中的一条恶意访问数据(例如:<script>alert('xss');</script>)对机器学习产品进行攻击访问时，预测判断模块将异常的恶意访问数据放入了正常访问特征值库中；

[0035] 恶意访问模块，用于采用一条恶意访问数据对机器学习产品进行攻击访问，该恶意访问数据中包括待预测的特征因子；

[0036] 预测结果模块，用于接收预测判断模块的预测输出结果，并判断恶意访问模块中的恶意访问数据对机器学习产品进行攻击访问的测试是否成功；在本实施例中，预测结果模块监控预测判断模块的预测输出为正常访问数据、此恶意数据可以以正常数据的身份进行访问，成功的绕过了机器学习产品的WAF(Web Application Firewall)防火墙系统，并弹出XSS的弹框，攻击者攻击成功。通过该系统对机器学习产品的测试结果说明该机器学习产品存在漏洞。

[0037] 以上所述实施例仅表达了本发明的具体实施方式，其描述较为具体和详细，但并不能因此而理解为对本发明专利范围的限制。应当指出的是，对于本领域的普通技术人员来说，在不脱离本发明构思的前提下，还可以做出若干变形和改进，这些都属于本发明的保护范围。

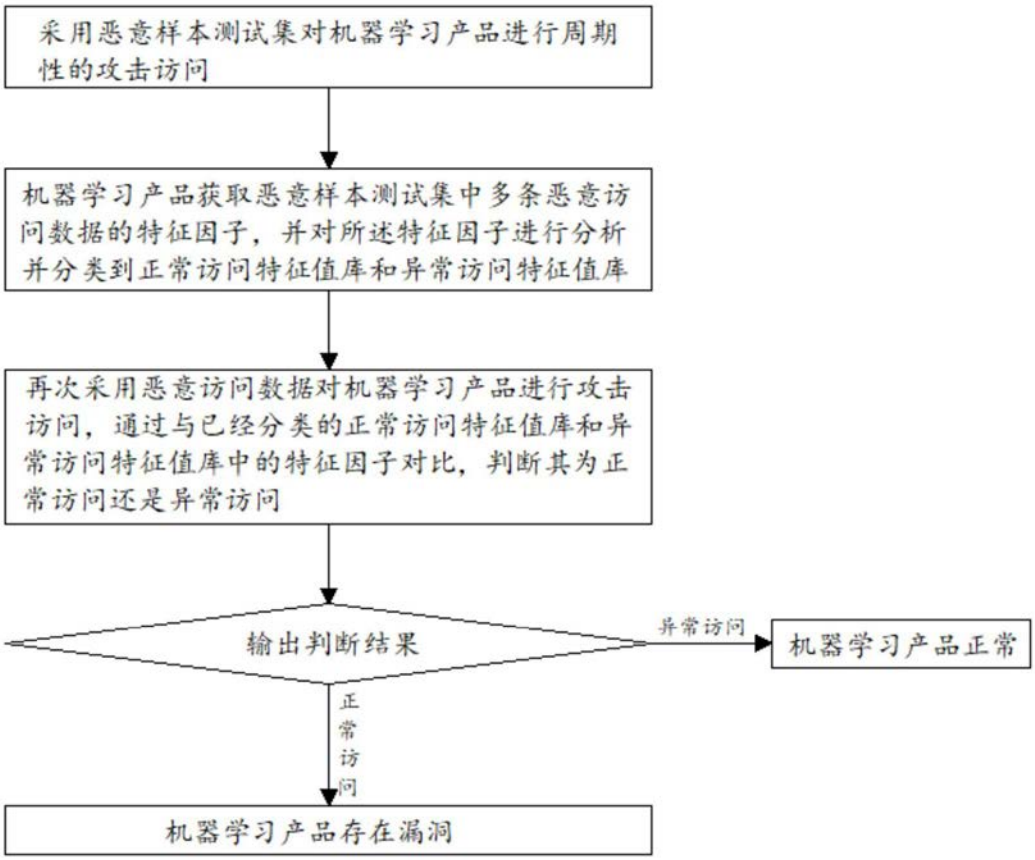

采用恶意样本测试集对机器学习产品进行周期性的攻击访问
机器学习产品获取恶意样本测试集中多条恶意访问数据的特征因子，并对所述特征因子进行分析并分类到正常访问特征值库和异常访问特征值库
再次采用恶意访问数据对机器学习产品进行攻击访问，通过与已经分类的正常访问特征值库和异常访问特征值库中的特征因子对比，判断其为正常访问还是异常访问
输出判断结果
异常访问
机器学习产品正常
正常访问
机器学习产品存在漏洞

图1

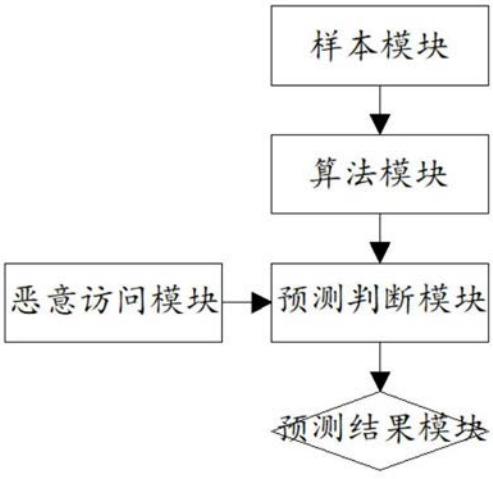

样本模块
算法模块
恶意访问模块
预测判断模块
预测结果模块

图2