(19)国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 118540076 A
(43)申请公布日 2024.08.23

(21)申请号 202310166120.8

(22)申请日 2023.02.22

(71)申请人 中国科学院计算机网络信息中心
地址 100190 北京市海淀区中关村南四街4号院内2号楼

(72)发明人 龙春 付豪 魏金侠 宫良一 付豫豪 王跃达

(74)专利代理机构 北京知舟专利事务所(普通合伙) 11550
专利代理师 周玉玲

(51)Int.Cl.
H04L 9/40(2022.01)
H04L 61/4511(2022.01)
G06F 18/214(2023.01)
G06F 18/2413(2023.01)
G06F 18/2431(2023.01)
G06N 20/20(2019.01)

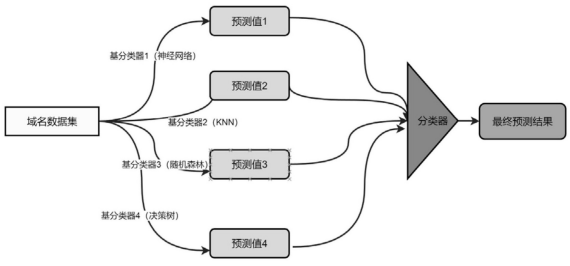
权利要求书1页 说明书5页 附图1页

(54)发明名称

基于多类特征的恶意域名识别方法

(57)摘要

本发明属于恶意域名检测技术领域,为解决现有技术对新的恶意域名识别率较低的技术问题,本发明提供一种基于多类特征的恶意域名识别方法,从包括恶意域名与正常域名的域名数据集中提取多类特征,并将同类特征组合成相应的训练集;多类特征包括IP通信流特征,IP通信流特征反应了域名的通信行为;为各类特征匹配相应的基分类器,并采用相应的训练集训练各个基分类器;采用训练完成后的各个基分类器对相应的训练集进行预测,并输出预测结果;将各个基分类器的预测结果组成新训练集,训练元分类器;将训练完成的基分类器与训练完成的元分类器集成为集成学习分类器,采用集成学习分类器对待测域名进行分类识别。本发明提高了识别率,降低了训练成本。



CN 118540076 A

1.一种基于多类特征的恶意域名识别方法,其特征在于,包括如下步骤:

从包括恶意域名与正常域名的域名数据集中提取多类特征,并将同类特征组合成相应的训练集;所述多类特征包括IP通信流特征,所述IP通信流特征反应了域名的通信行为;

为各类特征匹配相应的基分类器,并采用相应的训练集训练各个基分类器;

采用训练完成后的各个基分类器对相应的训练集进行预测,并输出预测结果;

将各个基分类器的预测结果组成新训练集,采用所述新训练集训练元分类器;

将训练完成的基分类器与训练完成的元分类器集成为集成学习分类器,采用所述集成学习分类器对待测域名进行分类识别。

2.根据权利要求1所述的基于多类特征的恶意域名识别方法,其特征在于,按如下方式提取IP通信流特征:

提取客户端与服务器之间的原始通信数据包,并保存为数据报存储格式;

对原始通信数据包进行分组处理,通过分组处理将解析同一域名数据获得的IP地址的TCP数据包划分为同一流量组;

对同一流量组进性会话还原后,提取包括流的持续时间、数据包大小、包间隔时间与端口号在内的流相关特征。

3.根据权利要求2所述的基于多类特征的恶意域名识别方法,其特征在于,会话还原包括对流量组中的TCP流量包进行合与按时间排序操作。

4.根据权利要求2所述的基于多类特征的恶意域名识别方法,其特征在于,所述流的持续时间按如下方式提取:计算流量组中第一个TCP数据包和最后一个TCP数据包的时间间隔;

所述数据包大小按如下方式提取:计算流量组中TCP数据包大小的期望、标准差以及中位数值;

所述包间隔时间按如下方式提取:计算流量组中相邻TCP数据包之间间隔时间的期望,标准差以及中位数值;

端口号按如下方式提取:提取流量组中各个TCP数据包端口号的集合。

5.根据权利要求2所述的基于多类特征的恶意域名识别方法,其特征在于,所述数据报存储格式为Pcap格式。

6.根据权利要求1所述的基于多类特征的恶意域名识别方法,其特征在于,所述多类特征还包括字符特征、注册相关特征与网络相关特征。

7.根据权利要求6所述的基于多类特征的恶意域名识别方法,其特征在于,为各类特征与基分类器的匹配关系如下:字符特征对应人工神经网络,注册相关特征对应k近邻分类器,网络相关特征对应随机森林分类器,IP通信流特征对应决策树。

8.根据权利要求1所述的基于多类特征的恶意域名识别方法,其特征在于,采用LR逻辑回归分类器作为所述元分类器。

9.根据权利要求1或8所述的基于多类特征的恶意域名识别方法,其特征在于,采用stacking集成学习方法训练所述元分类器。

10.根据权利要求1所述的基于多类特征的恶意域名识别方法,其特征在于,采用5折交叉验证方法训练各个基分类器。

基于多类特征的恶意域名识别方法

技术领域

[0001] 本发明属于恶意域名检测技术领域。

背景技术

[0002] 域名系统协议是互联网的重要组成部分,它将难以记忆的互联网协议地址映射到易于记忆的域名。大量的网络服务依赖于域名服务来展开。由于域名系统并不对依托于其开展的服务行为进行检测,DNS服务被滥用于各种恶意活动:传播恶意软件、促进命令和控制服务器通信、发送垃圾邮件、托管诈骗和网络钓鱼网页等。在攻击者的恶意行为中发挥了至关重要的作用,因此对恶意域名进行检测具有极其重要的意义。

[0003] 现有技术的缺点:1.大多数研究是基于域名字符特征的规律来识别合法域名与伪域名。由于新的恶意域名家族在不断涌现,特别是目前对由英语单词拼接的域名检测上效果不佳。由于不同恶意域名家族生成的域名数量不一,存在着训练数据过少、识别率低的问题。2.基于域名自身特性以及域名历史数据的方法难以识别出生存时间较少的恶意域名,且对新恶意域名检测效果较弱。

发明内容

[0004] 本发明的目的在于解决上述现有技术中存在的难题,提供一种基于多维特征的恶意域名识别方法,解决现有技术对新的恶意域名识别率较低的技术问题。

[0005] 本发明是通过以下技术方案实现的:一种基于多类特征的恶意域名识别方法,包括如下步骤:

[0006] 从包括恶意域名与正常域名的域名数据集中提取多类特征,并将同类特征组合成相应的训练集;所述多类特征包括IP通信流特征,所述IP通信流特征反应了域名的通信行为;

[0007] 为各类特征匹配相应的基分类器,并采用相应的训练集训练各个基分类器;

[0008] 采用训练完成后的各个基分类器对相应的训练集进行预测,并输出预测结果;

[0009] 将各个基分类器的预测结果组成新训练集,采用所述新训练集训练元分类器;

[0010] 将训练完成的基分类器与训练完成的元分类器集成为集成学习分类器,采用所述集成学习分类器对待测域名进行分类识别。

[0011] 进一步的,按如下方式提取IP通信流特征:

[0012] 提取客户端与服务器之间的原始通信数据包,并保存为数据报存储格式;

[0013] 对原始通信数据包进行分组处理,通过分组处理将解析同一域名数据获得的IP地址的TCP数据包划分为同一流量组;

[0014] 对同一流量组进性会话还原后,提取包括流的持续时间、数据包大小、包间隔时间与端口号在内的流相关特征。

[0015] 进一步的,会话还原包括对流量组中的TCP流量包进行合与按时间排序操作。

[0016] 进一步的,所述流的持续时间按如下方式提取:计算流量组中第一个TCP数据包和

最后一个TCP数据包的时间间隔；

[0017] 所述数据包大小按如下方式提取：计算流量组中TCP数据包大小的期望、标准差以及中位数值；

[0018] 所述包间隔时间按如下方式提取：计算流量组中相邻TCP数据包之间间隔时间的期望，标准差以及中位数值；

[0019] 端口号按如下方式提取：提取流量组中各个TCP数据包端口号的集合。

[0020] 进一步的，所述数据报存储格式为Pcap格式。

[0021] 进一步的，所述多类特征还包括字符特征、注册相关特征与网络相关特征。

[0022] 进一步的，为各类特征与基分类器的匹配关系如下：字符特征对应人工神经网络，注册相关特征对应k近邻分类器，网络相关特征对应随机森林分类器，IP通信流特征对应决策树。

[0023] 进一步的，采用LR逻辑回归分类器作为所述元分类器。

[0024] 进一步的，采用stacking集成学习方法训练所述元分类器。

[0025] 进一步的，采用5折交叉验证方法训练各个基分类器。

[0026] 与现有技术相比，本发明的有益效果包括：

[0027] 1.通过提取IP通信流特征(动态特征)可以更好地刻画域名的通信行为，域名的通信行为能够印证域名是否确实存在攻击行为，而不是像现有技术那样仅仅通过分析域名表层特征(静态特征)来判断，从而克服新的恶意域名家族识别率低的问题，同时具有更好的可解释性。

[0028] 2.本发明为每类特征匹配了适合各自的基分类器，故每个基分类器无需全部特征，训练成本较低，此外在实际情况下运行时可以根据数据特点和服务质量的需求增加或者删除基分类器。

[0029] 3.本发明通过科技网实验测试挑选出与IP通信流特征组合后检测效果最佳的特征组合：IP通信流特征、字符特征、注册相关特征与网络相关特征。并且这些特征能够在短时间内获取，适用于网关实时入侵检测系统。

[0030] 4.本发明采用Stacking集成学习方法将不同的异质基分类器结合在一起，发挥每个基分类器法的优势来获得更佳的预测性能。

[0031] 5.LR逻辑回归分类器与stacking集成方法的结合能够避免过拟合：采用加权的方法获取各个基分类器的预测结果，逻辑回归相当于给每个基分类器预测结果加权最后获得总的预测结果，也就是为了学习各个基分类器的重要程度，而不对基分类器结果本身进行过多运算。因此可以避免对基分类器结果的过拟合。

附图说明

[0032] 图1为集成学习分类器进行恶意域名分类的原理示意图。

具体实施方式

[0033] 传统的恶意域名检测方法只能通过域名本身的特征来检测，如果遇到新的恶意域名家族它的特征和已有的域名特征不同，已有的恶意域名构建的分类器将很难识别。本发明通过域名解析的IP来分析恶意行为也就是IP通信流特征，IP通信流特征是通过分析域名

有没有真正被用来进行攻击来检测恶意的,因此具有更好的检测效果。

[0034] 本发明所提出的IP通信流特征是一种动态特征,能够从数据流的角度来反映恶意域名所管理的流的行为特征,描述了域名解析的IP存在哪些活动。域名通过解析IP来进行通信,通过分析IP通信流特征可以印证域名确实存在攻击行为而不是仅仅通过分析域名表层特征来判断。

[0035] 本发明主要分为3个步骤:特征提取、恶意域名分类识别训练,采用训练完成的集成学习分类器进行恶意域名分类识别,下面结合附图对本发明作进一步详细描述。

[0036] 一)、特征提取

[0037] 从包括恶意域名与正常域名的域名数据集中提取多类特征,并将同类特征组合成相应的训练集;所述多类特征包括IP通信流特征,所述IP通信流特征反应了域名的通信行为。

[0038] 按如下方式提取IP通信流特征:

[0039] 提取客户端与服务器之间的原始通信数据包,并保存为数据报存储格式,如Pcap格式;

[0040] 对原始通信数据包进行分组处理,通过分组处理将解析同一域名数据获得的IP地址的TCP数据包划分为同一流量组;

[0041] 对同一流量组进性会话还原后,提取包括流的持续时间、数据包大小、包间隔时间与端口号在内的流相关特征。首先这三种特征容易计算获得,在实际场景下效率高。其次,端口代表计算机特定服务,反映了通信方的意图;持续时间和间隔时间反映了通信方网络包收发的频率;包大小反映了通信方具体通信的内容。虽然这些都是粗粒度的特征,但是这些特征可以简单有效地描述通信行为。在良性和大规模攻击(如ddos,僵尸网络)流量之间也会产生区别。此外结合此类特征也可以让整个模型可解释性更好。

[0042] 所述流的持续时间按如下方式提取:计算流量组中第一个TCP数据包和最后一个TCP数据包的时间间隔;

[0043] 所述数据包大小按如下方式提取:计算流量组中TCP数据包大小的期望、标准差以及中位数值;

[0044] 所述包间隔时间按如下方式提取:计算流量组中相邻TCP数据包之间间隔时间的期望,标准差以及中位数值

[0045] 端口号按如下方式提取:提取流量组中各个TCP数据包端口号并组成端口号集合,采用onehot编码方式。这里要提取出一组流量数据中的端口组成一个集合。对端口号集合采用onehot编码。优点:简单快捷。因为端口号本身没有意义,端口号代表使用了某类的服务,比如端口80或端口443可能是http相关服务。并且数字之间没有距离的概念,比如80和443数字本身差不能体现出代表意义的差别。所以采用代表类别的onehot编码。

[0046] 为了进一步提高检测精度,多类特征还包括字符特征、注册相关特征与网络相关特征,这些特性的提取方式属于现有技术,在此不再赘述。

[0047] (1)域名的字符特征,包括:

[0048] a)域名长度:域名字符串去除了顶级域名的字符串长度。在计算完所有域名长度后对其进行归一化处理,计算每个域名长度对最大域名长度占比。

[0049] b)域名中数字占比:域名中数字个数对域名长度的占比

[0050] c)随机性:计算域名中每个字符出现频率并以此计算信息熵

[0051] d)最长有意义字符串长度占比:计算域名中连续的最长单词对长度于整个域名字符串长度占比

[0052] e)域名中的三连词:通过自然语言处理技术提取出3-gram,即以长度为3在字符串上进行滑动窗口的操作,提取出每个窗口的字符串添加到字典中。最后对字典进行one-hot编码

[0053] f)到已知恶意域名的编辑距离:编辑距离即计算一个字符串变化到另一个字符串编辑的次数,有编辑操作为改变,增加和删除三种。之后计算每个编辑距离对最大编辑距离的占比。

[0054] (2)注册相关特征:攻击者往往会批量注册恶意域名来逃逸黑名单的检测,因此恶意域名在注册信息熵具有相关特征。

[0055] a)注册国家代码:域名注册地国家代码。

[0056] b)域名注册商:使用域名注册的商家的名称,如:Alibaba Cloud Computing Ltd

[0057] c)域名注册个体:域名注册个体或者企业的名称

[0058] d)域名的服务器:用来解析域名的权威名称服务器

[0059] e)Whois服务器信息:域名注册使用的whois服务器

[0060] f)注册日期:域名注册的日期

[0061] g)到期日期:域名到期的日期

[0062] 其中a到e类别信息采用onehot编码方式,f和g日期类型进一步计算出年、月、日、时、分、秒六个特征

[0063] (3)网络相关的特征:良性域名会经常更新ip等网络信息来保证服务的可用性,恶意域名提供的网络信息相对较少。

[0064] a)ip:域名反解析出来的ip数量,ip的变更次数以及ip是否出现在黑名单中

[0065] b)soa信息:具体包括了域名的刷新时间,重新请求时间,域名请求的TTL等。其中时间类型进一步计算出年、月、日、时、分、秒六个特征,TTL信息计算所在[0,1),[1,10),[10,100),[100,300),[300,900)的范围,TTL改变的次数,不同TTL值的个数,TTL记录的平均值。

[0066] c)cname信息:需要计算别名的数量

[0067] d)mx信息:包括域名的邮件交换服务器地址等,计算该地址是否出现在黑名单中

[0068] 二)、恶意域名分类识别训练

[0069] 为各类特征匹配相应的基分类器,并采用相应的训练集训练各个基分类器;采用训练完成后的各个基分类器对相应的训练集进行预测,并输出预测结果;将各个基分类器的预测结果组成新训练集,采用所述新训练集训练元分类器。

[0070] 采用集成学习的方法,通过多个基分类器模型组成的一个整体模型。集成学习主要分为bagging,boosting和stacking,本发明使用了stacking集成学习方法。主要过程为将训练好的所有基分类器对整个训练集进行预测,第j个基分类器对第i个训练样本的预测值将作为新训练集中第i个样本的第j个特征值。比如,对于每个基分类器每个batch大小为n,类别数为2,那么基分类器产生n*2的矩阵,基分类器有4个模型,会合并产生4*n*2大小的特征。

[0071] 最后对基于新的训练集进行训练，同时预测的过程也要先经过所有基分类器的预测形成测试集，最后再对测试集进行预测以验证训练效果。

[0072] 使用了4个基分类器，每个基分类器都使用除去测试集的训练集全部的数据进行训练。第一个基分类器使用人工神经网络，选取的特征为字符特征，神经网络学习能力强可以学习到字符串中不容易人工判别的特征。第二个模型使用k近邻分类器，选择注册相关的特征，KNN通过计算不同特征值之间的距离进行分类可以度量恶意域名之间的相似性，算法实现高效，同时加入新数据时可以不用重新训练。第三个模型使用随机森林分类器，选择网络相关的特征，由于网络数据常常存在缺失值，随机森林的方法具有较强的抗干扰能力，可以处理在存在特征遗失的情况下仍然维持较高的精度。第四个模型使用C4.5决策树，选择流相关的特征，决策树的方法较为容易理解和解释。最后将基分类器的结果再通过LR逻辑回归分类器进行预测得到最终的结果。

[0073] 在训练时将域名数据进行划分，采用5折交叉验证的方法，其中4份作为训练集，剩余1份作为验证集。训练集通过每个基分类器的预测结果形成新的训练集再通过LR逻辑回归分类器得到最后的结果。

[0074] 三)、恶意域名分类识别

[0075] 将训练完成的基分类器与训练完成的元分类器集成为集成学习分类器，采用所述集成学习分类器对待测域名进行分类识别。

[0076] 模型可以离线训练完成后在线部署实现实时检测，域名数据输入检测系统后可以获得结果，同时根据新的域名数据对模型进行定期更新。

[0077] 上述技术方案只是本发明的一种实施方式，对于本领域内的技术人员而言，在本发明公开了原理的基础上，很容易做出各种类型的改进或变形，而不仅限于本发明上述具体实施例所描述的技术方案，因此前面描述的只是优选的，而并不具有限制性的意义。

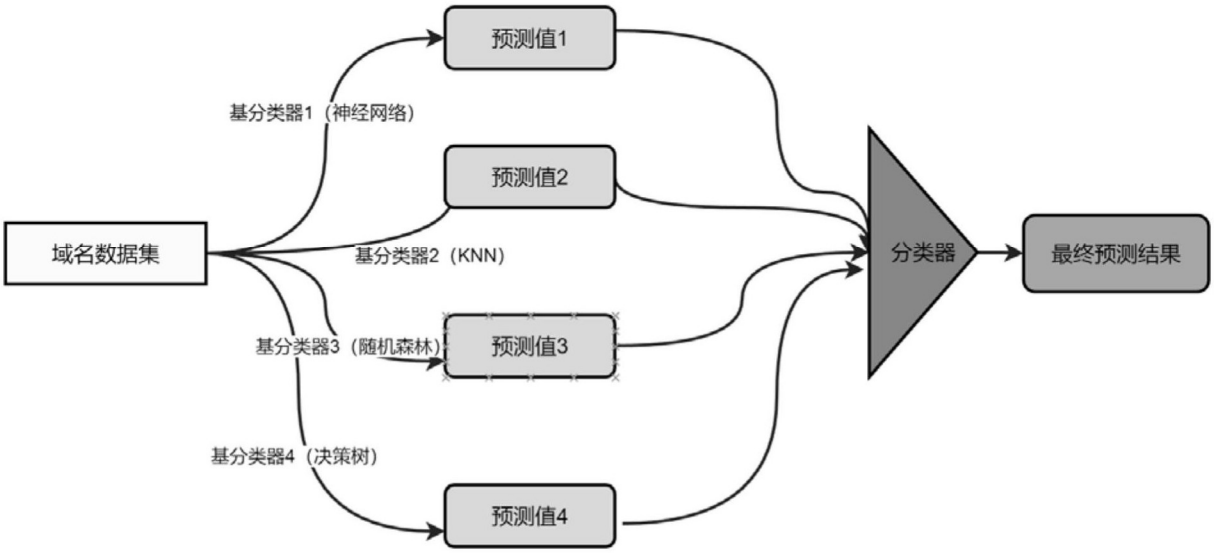

预测值1
基分类器1（神经网络）
预测值2
域名数据集
基分类器2（KNN）
分类器
最终预测结果
基分类器3（随机森林）
预测值3
基分类器4（决策树）
预测值4

图1