(19)中华人民共和国国家知识产权局

(12)发明专利

(10)授权公告号 CN 111464510 B
(45)授权公告日 2021.06.08

(21)申请号 202010191536.1

(22)申请日 2020.03.18

(65)同一申请的已公布的文献号
申请公布号 CN 111464510 A

(43)申请公布日 2020.07.28

(73)专利权人 华南理工大学
地址 510640 广东省广州市天河区五山路381号

(72)发明人 金冬子 陆以勤 覃健诚 王君君 毛中书 李佳

(74)专利代理机构 广州市华学知识产权代理有限公司 44245
代理人 李斌

(51)Int.Cl.
H04L 29/06(2006.01)
G06K 9/62(2006.01)

(56)对比文件
CN 110138786 A,2019.08.16
CN 110378430 A,2019.10.25
CN 110868409 A,2020.03.06
HAYEL KHAFAJEH.AN EFFICIENT INTRUSION DETECTION APPROACH USING LIGHT GRADIENT BOOSTING.《Journal of Theoretical and Applied Information Technology》.2020,第98卷(第05期),第3节第1段至第5节最后一段.

审查员 薛乐梅

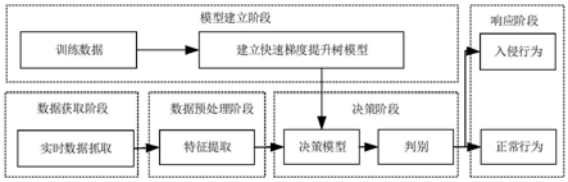
权利要求书2页 说明书4页 附图1页

(54)发明名称

基于快速梯度提升树分类模型的网络实时入侵检测方法

(57)摘要

本发明公开了一种基于快速梯度提升树模型的网络实时入侵检测方法,该方法包括下述步骤:使用训练数据训练快速梯度提升树分类模型;在连续的时间窗口抓取网络流量数据,每个时间窗口中抓取的原始流量数据作为一个数据块;对数据块进行统计分析,生成多个特征向量;采用快速梯度提升树分类模型对特征向量进行分类,区分正常行为和网络入侵行为;若判定为网络入侵行为,输出网络入侵告警信号。本发明能够克服分类性能的类别偏向性问题和降低误报率,同时模型建立阶段和决策阶段的匹配过程满足实时性的要求。



CN 111464510 B

1.一种基于快速梯度提升树分类模型的网络实时入侵检测方法,其特征在于,包括下述步骤:

使用训练数据训练快速梯度提升树分类模型;

所述快速梯度提升树分类模型的具体训练步骤包括:

所述快速梯度提升树分类模型由M棵决策树构成,初始化第一棵决策树为常数,表示为:

$$\hat{y}_i^{(0)} = f_0 = 0$$

其中,f_0代表初始化的决策树,$\hat{y}_i^{(0)}$代表初始化的预测值;

训练下一棵决策树,使用按叶子的生长策略,通过最小化损失函数得到第t次迭代中得到的决策树模型,表示为:

$$f_t(x_i) = \arg\min_{f_t} L_{(t)} = \arg\min_{f_t} L\left(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)\right)$$

其中,$f_t(x_i)$表示在第t次迭代中得到的决策树模型,$L_{(t)}$表示损失函数,y_i表示第i个实例的标签值,$\hat{y}_i^{(t)}$表示在第t次迭代中对第i个实例的预测值;

将上一次迭代中的决策树模型与当前次迭代中得到的决策树模型相加,得到新的模型:

$$\hat{y}_i^{(t)} = \hat{y}_i^{(t-1)} + f_t(x_i)$$

完成M次模型迭代后,得到并返回最终的训练模型:

$$\hat{y}_i^{(t)} = \sum_{t=0}^{M-1} f_t(x_i)$$

所述训练数据的具体构建步骤包括:

对原始数据采用GOSS进行采样,采用EFB对互斥的稀疏特征绑定后得到训练数据,表示为特征向量集:

$$\{(x_i, y_i)\}_{i=1}^{N}$$

其中,N表示特征向量的个数,x_i表示特征属性,y_i表示标签;

在连续的时间窗口抓取网络流量数据,每个时间窗口中抓取的原始流量数据作为一个数据块;

对所述数据块进行统计分析,生成多个特征向量;

采用所述快速梯度提升树分类模型对所述特征向量进行分类,区分正常行为和网络入侵行为;

若判定为网络入侵行为,输出网络入侵告警信号。

2.根据权利要求1所述的基于快速梯度提升树分类模型的网络实时入侵检测方法,其特征在于,所述在连续的时间窗口抓取网络流量数据,具体采用Tcpdump在连续的时间窗口内对主机网卡进行监听,抓取网络中的原始流量数据。

3.根据权利要求1所述的基于快速梯度提升树分类模型的网络实时入侵检测方法,其特征在于,所述生成多个特征向量,具体步骤包括:

根据源IP和目的IP将所述数据块中的数据划分成双向流；

采用CICFlowMeter对数据块进行基于双向流的特征提取，生成特征向量。

4.根据权利要求1所述的基于快速梯度提升树分类模型的网络实时入侵检测方法，其特征在于，所述区分正常行为和网络入侵行为，具体步骤包括：

将特征向量代入快速梯度提升树分类模型进行判别，得到预测值 $\hat{y}_i$ ；

若预测值 $\hat{y}_i$ 与正常行为的标签相同，则判定当前流属于正常行为；若与入侵类别标签相同，则判定当前流属于入侵行为。

5.根据权利要求1所述的基于快速梯度提升树分类模型的网络实时入侵检测方法，其特征在于，所述网络入侵告警信号包括网络入侵行为发生时间、网络入侵行为类别信息和网络入侵行为的网络来源。

基于快速梯度提升树分类模型的网络实时入侵检测方法

技术领域

[0001] 本发明涉及网络安全领域，具体涉及一种基于快速梯度提升树分类模型的网络实时入侵检测方法。

背景技术

[0002] 常见的网络安全防护技术有防火墙，数据加密，认证和数字签名等，这些静态的被动防守式网络安全工具服务方式单一，难以应对当今复杂多变的网络入侵。

[0003] 网络流量是入侵检测的一个重要数据来源，常见的网络入侵检测方法可以根据原理分为基于误用的网络入侵检测和基于异常的网络入侵检测，但现实网络环境中得到的训练数据为不平衡数据，因此基于误用的入侵检测方法通常在分类性能上存在类别偏向性，即对于一些数据量少的入侵行为检测性能较差，此外，基于误用的入侵检测方法常常比较复杂，因而难以满足实时性的要求；另一方面，界定正常行为和入侵行为本就是具有挑战性的任务，况且当前环境中正常用户的行为并非静态不变的，基于异常的网络入侵检测可能会把正常行为误判为入侵行为，因而常具有较高的误报率。

发明内容

[0004] 为了克服现有技术存在的缺陷与不足，本发明提供一种基于快速梯度提升树分类模型的网络实时入侵检测方法，能够克服分类性能的类别偏向性问题和降低误报率，同时模型建立阶段和决策阶段的匹配过程满足实时性的要求。

[0005] 为了达到上述目的，本发明采用以下技术方案：

[0006] 本发明提供一种基于快速梯度提升树分类模型的网络实时入侵检测方法，包括下述步骤：

[0007] 使用训练数据训练快速梯度提升树分类模型；

[0008] 在连续的时间窗口抓取网络流量数据，每个时间窗口中抓取的原始流量数据作为一个数据块；

[0009] 对所述数据块进行统计分析，生成多个特征向量；

[0010] 采用所述快速梯度提升树分类模型对所述特征向量进行分类，区分正常行为和网络入侵行为；

[0011] 若判定为网络入侵行为，输出网络入侵告警信号。

[0012] 作为优选的技术方案，所述训练数据的具体构建步骤包括：

[0013] 对原始数据采用GOSS进行采样，采用EFB对互斥的稀疏特征绑定后得到训练数据，表示为特征向量集：

[0014] $\left\{\left(x_i, y_i\right)\right\}_{i=1}^{N}$

[0015] 其中，N表示特征向量的个数，x_i表示特征属性，y_i表示标签。

[0016] 作为优选的技术方案，所述快速梯度提升树分类模型的具体训练步骤包括：

[0017] 所述快速梯度提升树分类模型由M棵决策树构成，初始化第一棵决策树为常数，表示为：

[0018] $\hat{y}_i^{(0)} = f_0 = 0$

[0019] 其中，f_0代表初始化的决策树，$\hat{y}_i^{(0)}$代表初始化的预测值；

[0020] 训练下一棵决策树，使用按叶子的生长策略，通过最小化损失函数得到第t次迭代中得到的决策树模型，表示为：

[0021] $$f_t(x_i) = \arg\min_{f_t} L_{(t)} = \arg\min_{f_t} L\left(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)\right)$$

[0022] 其中，$f_t(x_i)$表示在第t次迭代中得到的决策树模型，$L_{(t)}$表示损失函数，y_i表示第i个实例的标签值，$\hat{y}_i^{(t)}$表示在第t次迭代中对第i个实例的预测值；

[0023] 将上一次迭代中的决策树模型与当前次迭代中得到的决策树模型相加，得到新的模型：

[0024] $$\hat{y}_i^{(t)} = \hat{y}_i^{(t-1)} + f_t(x_i)$$

[0025] 完成M次模型迭代后，得到并返回最终的训练模型：

[0026] $$\hat{y}_i^{(t)} = \sum_{t=0}^{M-1} f_t(x_i)$$ 。

[0027] 作为优选的技术方案，所述在连续的时间窗口抓取网络流量数据，具体采用Tcpdump在连续的时间窗口内对主机网卡进行监听，抓取网络中的原始流量数据。

[0028] 作为优选的技术方案，所述生成多个特征向量，具体步骤包括：

[0029] 根据源IP和目的IP将所述数据块中的数据划分成双向流；

[0030] 采用CICFlowMeter对数据块进行基于双向流的特征提取，生成特征向量。

[0031] 作为优选的技术方案，所述区分正常行为和网络入侵行为，具体步骤包括：

[0032] 将特征向量代入快速梯度提升树分类模型进行判别，得到预测值$\hat{y}_i$；

[0033] 若预测值$\hat{y}_i$与正常行为的标签相同，则判定当前流属于正常行为；若与入侵类别标签相同，则判定当前流属于入侵行为。

[0034] 作为优选的技术方案，所述网络入侵告警信号包括网络入侵行为发生时间、网络入侵行为类别信息和网络入侵行为的网络来源。

[0035] 本发明与现有技术相比，具有如下优点和有益效果：

[0036] (1)本发明基于快速梯度提升树模型进行网络入侵检测，在模型训练前先采用GOSS对样本数据采样以减低数据量，通过EFB对互斥的稀疏特征进行绑定以降低特征维度，进而到达到提升模型建立阶段实时性的目的。

[0037] (2)本发明的快速梯度提升树模型在生成每个决策树时，采用按叶生长的策略，在增长一个叶子节点的情况下，比按层生长的策略能够降低更多的误差；为了防止模型过拟合，快速梯度提升树模型限制每个决策树的深度，在最终得到的模型中由较少的决策树和叶子节点构成，这一特点使得快速梯度提升树模型在决策阶段的匹配过程具有良好的时间高效性。

[0038] (3)本发明在每次训练新的决策树时对原始样本数据采用GOSS方法进行采样，再

将决策树加起来得到最终的预测结果，一定程度上能够克服训练数据的不平衡，并且起到改善检测性能，特别是降低误报率的效果。

附图说明

[0039] 图1为本实施例基于快速梯度提升树分类模型的网络实时入侵检测方法的流程示意图；

[0040] 图2为本实施例建立基于快速梯度提升树的分类模型的流程示意图。

具体实施方式

[0041] 为了使本发明的目的、技术方案及优点更加清楚明白，以下结合附图及实施例，对本发明进行进一步详细说明。应当理解，此处所描述的具体实施例仅仅用以解释本发明，并不用于限定本发明。

[0042] 实施例

[0043] 如图1所示，本实施例提供一种基于快速梯度提升树分类模型的网络实时入侵检测方法，包括下述步骤：

[0044] S1、模型建立阶段，使用训练数据训练快速梯度提升树分类模型；

[0045] 如图2所示，建立基于快速梯度提升树的分类模型具体步骤包括：

[0046] 对原始数据采用GOSS(Gradient-based One-Side Sampling)进行采样从而降低数据量，同时通过EFB(Exclusive Feature Bundling)对互斥的稀疏特征绑定从而降低特征维度，之后得到的数据作为训练数据；

[0047] 训练数据集来源于入侵检测系统所部署的网络环境，表示为特征向量集：

[0048] $\{(x_i, y_i)\}_{i=1}^{N}$

[0049] 其中，N代表特征向量的个数，x_i为特征属性，y_i为标签；

[0050] 快速梯度提升树模型由M棵决策树构成，将第一棵决策树初始化为常数，其表示形式如下所示：

[0051] $\hat{y}_i^{(0)} = f_0 = 0$

[0052] 其中，f_0代表初始化的决策树，$\hat{y}_i^{(0)}$代表初始化的预测值；

[0053] 训练下一棵决策树，使用按叶子的生长策略，通过最小化损失函数得到，其表示形式如下：

[0054] $f_t(x_i) = \arg\min_{f_t} L_{(t)} = \arg\min_{f_t} L\left(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)\right)$

[0055] 其中y_i是第i个实例的标签值，$\hat{y}_i^{(t)}$是在第t次迭代中对第i个实例的预测值，$f_t(x_i)$代表在第t次迭代中得到的决策树模型，$L_{(t)}$是损失函数，用来衡量预测值$\hat{y}_i^{(t)}$和目标值y_i的误差；

[0056] 迭代得到新的模型，将上一次迭代中的模型与本次迭代中得到的决策树相加，得到本次迭代的模型，其表示形式如下：

[0057] $\hat{y}_i^{(t)} = \hat{y}_i^{(t-1)} + f_t(x_i)$；

[0058] 之后继续迭代，通过最小化损失函数生成新的决策树，并通过加法方式得到新一轮迭代中的模型；

[0059] 判定是否达到停止条件，迭代的停止条件为完成第M次迭代，其中M为预先确定的迭代次数，完成M次模型迭代后，得到并返回最终的训练模型，其表示形式如下：

[0060] $$\hat{y}_i^{(t)} = \sum_{t=0}^{M-1} f_t(x_i)$$

[0061] 至此完成基于快速梯度提升树的分类模型的建立；

[0062] S2、数据获取阶段，在连续的时间窗口从网络中抓取流量数据，每个时间窗口中抓取的原始流量数据作为一个数据块；

[0063] 具体步骤为：采用Tcpdump在连续的时间窗口内对主机网卡进行监听，抓取网络中的原始流量数据；

[0064] 单个时间窗口的长度取为2s，每个时间窗口中抓取的数据被存储为一个数据块，文件格式为.pcap；

[0065] S3、数据预处理阶段，通过对每一个数据块进行统计分析，生成若干个特征向量；

[0066] 具体步骤为：将数据块中的数据，根据源IP和目的IP将数据包划分成双向流；

[0067] 采用CICFlowMeter对数据块进行基于双向流的特征提取，生成由84个特征表示的特征向量；

[0068] S4、决策阶段，通过已经建立的快速梯度提升树模型对特征向量进行分类，分为正常行为和若干具体的入侵行为；

[0069] 具体步骤为：将特征向量代入基于快速梯度提升树的分类模型进行判别，得到预测值$\hat{y}_i$；

[0070] 若$\hat{y}_i$与正常行为的标签相同，则认为该流为正常行为，否则若与具体的入侵类别标签相同，认为该流属于这类入侵行为；

[0071] S5、响应阶段，若决策阶段发现入侵行为，入侵检测系统向系统管理员发送告警信号；

[0072] 具体步骤包括：对检测到的入侵行为，向系统管理员发送告警信号，该信号内容包括入侵行为发生时间、入侵行为的具体类别信息和入侵行为的网络来源。

[0073] 通过以上的实施方式的描述，本领域的技术人员可以清楚地了解到本发明可借助软件加必需的硬件平台的方式来实现，当然也可以全部通过硬件来实施，但很多情况下前者是更佳的实施方式。基于这样的理解，本发明的技术方案对背景技术做出贡献的全部或者部分可以以软件产品的形式体现出来，该计算机软件产品可以存储在存储介质中，如ROM/RAM、磁碟、光盘等，包括若干指令用以使得一台计算机设备(可以是个人计算机，服务器，或者网络设备等)执行本发明各个实施例或者实施例的某些部分所述的方法。

[0074] 上述实施例为本发明较佳的实施方式，但本发明的实施方式并不受上述实施例的限制，其他的任何未背离本发明的精神实质与原理下所作的改变、修饰、替代、组合、简化，均应为等效的置换方式，都包含在本发明的保护范围之内。

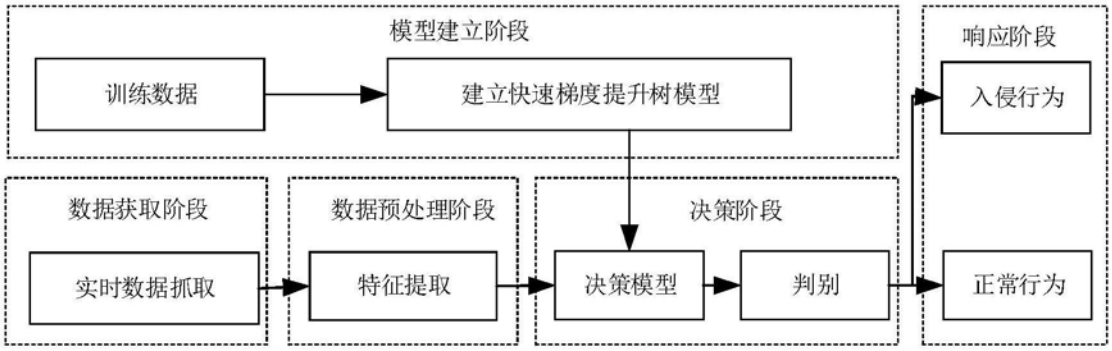

模型建立阶段
训练数据
建立快速梯度提升树模型
响应阶段
入侵行为
数据获取阶段
实时数据抓取
数据预处理阶段
特征提取
决策阶段
决策模型
判别
正常行为

图1

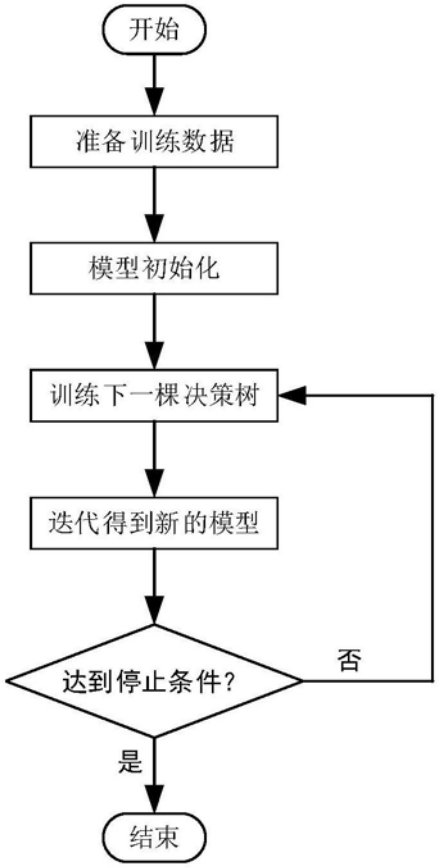

开始
准备训练数据
模型初始化
训练下一棵决策树
迭代得到新的模型
达到停止条件?
否
是
结束

图2