A Lightweight Traffic Anomaly Detection Model in SDN Based on Decision Tree

3rd International Conference on Communications, Information Management and Network Security (CIMNS 2018)

两篇引用论文

在相关研究中，论文[3]概述了如何使用SDN检测网络攻击，并得出结论，SDN是一个适合于抵御DDOS攻击的平台，主要是因为使用了标准协议、服务和接口，从而促进了新解决方案的部署。在实际网络中，流量异常往往是由DDOS和端口扫描攻击触发的。

五个特征

论文[4]提出了一种基于流量采集、特征提取和流量分类的分布式拒绝服务攻击检测解决方案。使用的流量特征是每个流量的平均包数、每个流量的平均字节数、每个流量的平均持续时间、成对流量的百分比以及单个流量的增长。基于这五个索引，使用自组织映射建立检测模型来执行流量异常检测。论文[5]提出了一种基于sflow的从网络中获取信息的机制，并利用信息论进行包分类，然后改变流规则来阻止恶意流。

决策树：

决策树是最有效的分类算法之一。

决策树的三个特点

特征选择措施：选择属性，对训练数据进行高精度的分割，评估每个属性对训练项目的分类能力。

划分策略：通过考虑所选测试属性的值，将当前训练集划分为几个子集。

停止条件：它结束树的一部分的生长，从而将训练子集声明为叶。

对于决策树，属性是离散的或连续的，在这里我们将特别关注连续属性，并使用多项式拟合来对网络流量进行分类。



拟合系数，…用于最小化训练数据的均方误差（mse）：



事实上，

决策树归纳算法的目标是最小化两个参数：

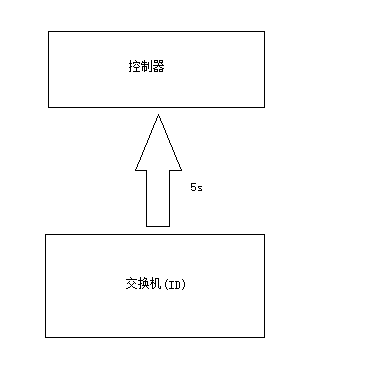
错误分类率和树大小。

误分类率：分类器的主要目标是追求最佳分类率。一个模型不仅要对训练数据进行正确的分类，而且要具有良好的泛化能力，才能对其他新的训练数据进行正确的分类。误分类率能有效地平衡一般性和准确性。

树大小：为了更好的泛化能力和更好的分类CPU时间和内存使用，大多数决策树学习算法倾向于构建小的树。为了建立这样的树，决策树归纳算法在树根附近建立具有最明显属性的决策树，以最小化区分不同类所需的测试数量。此外，决策树算法通常涉及到修剪阶段，这会消除子树，而子树不能确保显著的分类增益。另一种最小化树大小的机制是在诱导阶段停止树生长的停止标准之一。

流量异常检测模型

1. Collecting Flows Using Floodlight/OpenFlow



控制器以预定的时间间隔（本文中为5秒）从OF交换机收集流量入口。从这个集合中提取重要的特性，将流量分类为正常流量或攻击流量。当收集器从经过flootlight认证的所有交换机中收集样本时，交换机ID用于帮助分类器模块查明检测到哪些交换机的DDOS洪泛攻击。

1. Selecting Features

本文通过比较分析，构建了五个统计指标，建立了决策树模型。首先假设N表示流表中的流条目数，Pi表示流条目Fi的包数，表示流条目Fi的包Pi中的字节数。根据这些信息，这些特征的值可以计算如下：

（1）APF（每个流入口的数据包平均值）：APF是一个重要的网络流量监控索引。例如，在ddos、端口扫描攻击中，每个流条目的平均数据包数将显著减少或增加。在传统的网络中，我们可以通过网络流量等网络测量工具直接获得该指标的值。但网络流是基于采样的，它不能检测到低频攻击，会对设备的性能造成危害。



1. ABF（每个流条目的平均字节数）：ABF表示流表在一个时间窗口中处理的有效负载大小。当像ddos这样的攻击发生时，它通常非常小。



1. PPF（成对流入口的百分比）：成对流入口是指处理单向和反向数据包的两个流入口。例如，流条目x=（srcip=a，destip=b）和流条目y=（srcip=b，destip=a）是对流条目，ppf可以通过以下方式计算：



PF表示具有匹配对流入口的流入口数。PPF是SDN网络中的一个特殊功能。没有响应的大量数据包会导致PPF显著增加。

（4）SPF（流表匹配的源IP数量）：以连续时间为单位反映源IP号的变化，该指标可以测量流量中伪造的源IP地址，对欺骗攻击敏感。 它可以计算如下：



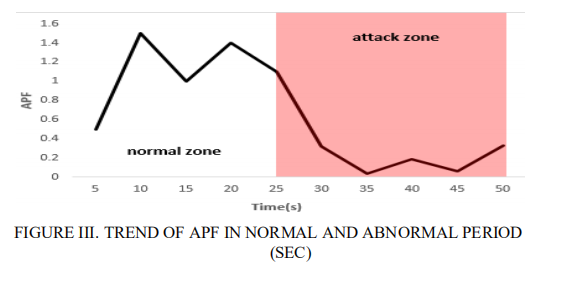
srcip\_num是给定时间间隔内单个流表匹配的数据包的源IP数。

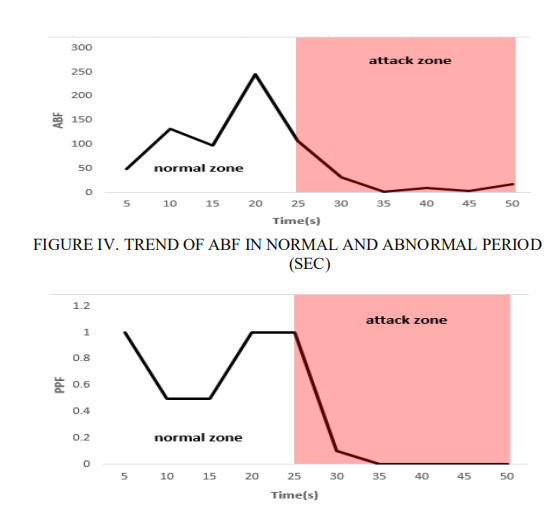
（5）ADF（每个流入口的平均持续时间）：每个流入口都有自己的生命周期。在OF交换机中，如果一个流入口在给定的时间段内没有匹配的数据包，系统将删除该数据包。如论文[4]所述，当应用程序之间交换的数据包数量很少时，此功能可以减少误报的数量。

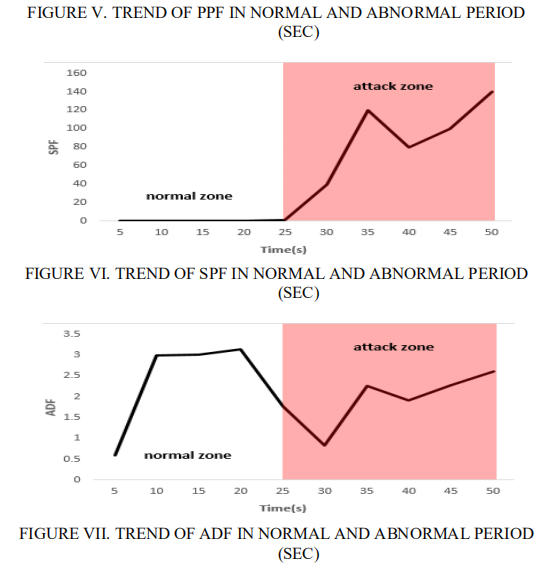


Di表示第i个流入口的持续时间。流量异常总是导致ADF下降。

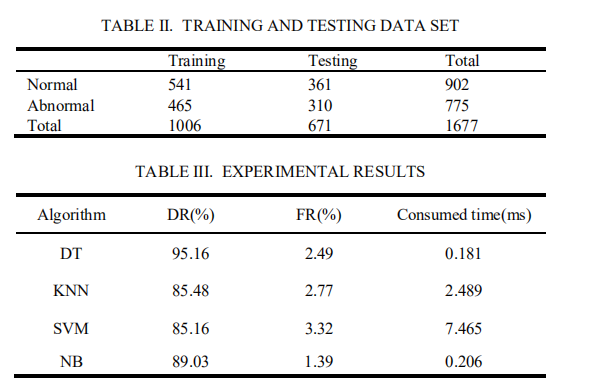
实验：







为了测试决策树算法和其他一些机器学习算法的差异，我们收集了正常和异常期间的1677个流条目进行培训和测试，如表2所示。



采用决策树、K最近邻、支持向量机和朴素贝叶斯算法对数据集进行训练和测试。四种算法的检测率、误报率和测试时间如表3所示。

实验结果表明，决策树的检测率最高，为95.16%，误报率为2.49%，测试时间最短，为0.181ms，虽然朴素贝叶斯的误报率最低，为1.39%，但其检测率明显低于决策树。K-最近邻和支持向量机的检测率和误报率也明显低于决策树。

通过比较四种算法的分类效果，可以看出决策树是检测网络中异常流量的最佳方法。