·CTCIS 2016 推荐论文· DOI:10.15961/j.jsuese.2017.01.019

基于多维时间序列分析的网络异常检测

陈兴蜀[1],江天宇[2],曾雪梅[1*],尹学渊[2],邵国林[2]
(1.四川大学 网络空间安全研究院,四川 成都 610065;2.四川大学 计算机学院,四川 成都 610065)

摘 要:针对实际网络异常检测要求高检测率、低误报率的问题,提出了一种基于多维时间序列的检测方法。首先,通过对实际网络流量进行长期观测,提取多维特征对网络流量进行描述;然后,利用时间序列分析方法对多维特征进行预测,计算预测值与真实值的时间序列偏离度,并且实时更新偏离度,适应多变的网络环境;最后,利用支持向量机(SVM)算法对偏离度向量进行分类判别,判断是否发生异常。目前该方法已应用于校园网关键服务器的实时监测与防护工作中,实际服务器流量的预测、告警结果表明,该方法可以有效检测网络中的异常流量。

关键词:异常检测;时间序列;网络流量;多维特征;网络安全
中图分类号:TP393.08 **文献标志码**:A **文章编号**:2096-3246(2017)01-0144-07

Network Anomaly Detector Based on Multiple Time Series Analysis

CHEN Xingshu[1],*JIANG Tianyu*[2],*ZENG Xuemei*[1*],*YIN Xueyuan*[2],*SHAO Guolin*[2]
(1. Cybersecurity Research Inst., Sichuan Univ., Chengdu 610065, China; 2. College of Computer Sci., Sichuan Univ., Chengdu 610065, China)

Abstract: The anomaly detection of network traffic in practice requires both high detection rate and low false alarm rate. To address this problem, a detection approach based on multidimensional time series analysis was proposed. Firstly, the network traffic was observed in a long time, and multiple network features were chosen for building the network behavior model. Subsequently, multiple features were predicted by the method of time series analysis. Then the degree of deviation between the predict value and the real value was calculated and updated. Finally, the state of whether the network flow is normal was determined by using support vector machine to classify the degree of deviation in time series. This method has been applied to real-time monitoring and protection on a campus key server. The results showed that it can detect anomalies effectively in network traffic.

Key words: anomaly detection; time series; network traffic; multiple features; network security

网络异常流量检测是网络安全防护的重要组成部分,也是目前学术界和产业界研究的热点,它主要是通过分析流经目标系统的所有网络流量来发现网络异常。根据检测思路的不同,网络异常检测主要可以分为以下两大类,基于误用的检测和基于异常的检测[1-2]。前者主要是对已知攻击手段提取特征码,然后检测当前网络流量中是否符合这些特征码,一旦匹配,则认为发生了异常,目前广泛使用的入侵检测系统(intrusion-detection system, IDS)、入侵防御系统(intrusion prevention system, IPS)就属于这类方法,这类方法的优点是检测准确率高,缺点就是一旦攻击者改变特征,容易绕过防御,漏报率高。而基于异常的检测的主要思路是为目标系统定义一个正常的行为模型,一旦目标系统偏离正常的行为模型,就判定为非法行为。这种方法的优点是通用性强,可以检测出未知异常。但是在实际应用中存在误报率高的情况。

在网络异常流量检测过程中,往往利用网络流量在时间序列上的变化情况来检测异常。文献[3]利用利用改进的 Holter-Winters 算法对网络流量进行预测,取得了一定的效果。但是由于其仅对网络流量特征进行检测,对于目前日益复杂的网络异常流量,往往难以检测。文献[4-5]在多个不同维度不同层次上的分布情况对网络流量进行描述,与单

收稿日期:2016-09-18
基金项目:国家自然科学基金资助项目(61272447)
作者简介:陈兴蜀(1968—),女,教授,博士生导师,博士. 研究方向:云计算;信息安全;计算机网络. E-mail:chenxsh@scu.edu.cn
*通信联系人 E-mail:zengxm@scu.edu.cn

一维度相比,检测精度有较大提高。文献[6]提出了一种基于数据流结构稳定性(FSS)的检测算法,利用 AR 自回归模型估计 FSS 时间序列多维特征,最后利用 SVM 进行异常判别,该方法对于 DDoS 具有一定的检测效果。文献[7]提取网络流量中多个流量属性的概率分布时间序列表示为多维信息散度向量,然后建立自回归滑动平均(ARMA)检测该向量是否异常,该方法对于僵尸网络具有一定的检测效果。文献[8]则将时间序列分析方法用于工业控制以太网的流量异常检测,与传统的时间序列分析相比,该方法对于异常检测系统的效率有很大改进。文献[9]利用 ARIMA 算法对 web 服务中的正常行为建立模型,当特征值超过正常行为的置信区间时,则判定为异常,该方法对于实际数据中的异常检测具有较好的效果。文献[10]利用自回归模型拟合得到网络数据流量的多维参数向量,以此描述单位时间内网络数据流量势能的稳定性,最终利用支持向量机对网络流量特征参数进行分类。上述研究方法大多数都是针对网络中 5 元组信息、流量信息进行检测,并不能完全反应网络实际状况,检测异常类型受到很大限制,检测精度也偏低。

本文首先分析了网络实际流量特征,提取 TCP 会话过程中多个特征维度的信息,定义了时间序列偏离度,并介绍了偏离度的更新算法,把不同维度上的偏离度排列成多维偏离度检测向量,然后利用支持向量机(support vector machine,SVM)进行分类。通过对实际网络流量进行训练,最终检测出真实网络环境中的异常流量。

本文的创新点主要为:

1)通过对实际网络流量进行长期观测,提取出能够描述正常网络行为的多维特征,在各个维度上进行分析,提高了异常检测的准确率。

2)在时间序列分析的基础上引入了时间序列偏离度的概念,并且对偏离度进行更新,能够适应复杂多变的网络环境。

3)利用分类算法,能够结合各个维度的时间序列偏离度进行综合判断,提高了异常检测的准确率。

1　多维特征分析及提取

通过对校园环境内实际流量统计分析,TCP 流量占了整体流量的大部分,而且很多攻击也是针对 TCP 的,因此本文只考虑 TCP 流量的特征提取。

目前,网络异常检测的特征主要基于数据包级别和会话流级别。数据包级别的特征主要是数据包长度,数据包数量等。TCP 会话流是指从主机发送 SYN,3 次握手建立连接到 4 次挥手连接结束的过程,在实际情况中也会存在连接超时等情况,设定超时时间为 T。一旦会话流中在时间 T 内没有任何数据包,则会话流结束。本文将会从这两类特征中进行特征选择。文献[11]提出了用于刻画一个完整的 TCP 流的 248 个特征,文献[12]通过分析用户异常行为对流量统计的影响,提出了一个较为完备的网络流量特征集,包括包长类、地址类、端口类、速度类、分布类等网络流量特征。但这些特征如果全部采集,会降低程序效率,而且部分特征对异常检测并无贡献。因此,本文对这些特征进行了筛选和拓展。

定义 1(主动连接)　根据 TCP 连接中的源 IP、目的 IP、源端口、目的端口,4 元组标志一条 TCP 连接,当主机为连接发起方(发送第一个 SYN 请求)时,则为主动连接。

定义 2(被动连接)　根据 TCP 连接中的源 IP、目的 IP、源端口、目的端口,4 元组标志一条 TCP 连接,当主机不是连接发起方(发送第一个 SYN 请求)时,则为被动连接。

对于主机主动发起的连接,可以描绘主机与外界通信的频繁程度,主动连接一般反应的是作为客户端的特征。如果网络中存在频繁的主动连接,有可能是僵尸主机,正在向外界发动 DDoS 攻击等。对于服务端在周知端口(如 80、25 等端口)上的监听,可以有效地描绘出服务器网络业务频繁程度、业务质量等问题。

在每个统计窗口 t 内,对特征量 F 进行统计,在连续 N 个时间窗口内,就可以得到时间序列 $F_1, F_2, F_3, \cdots, F_N$。

图 1 ~4 分别是某服务器被动连接中,上行流量、下行流量、通信 IP 个数、TCP 连接数随时间变化图。

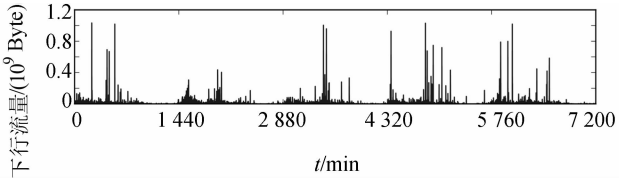

图 1　上行流量时间序列

Fig. 1　Upstream traffic in time series

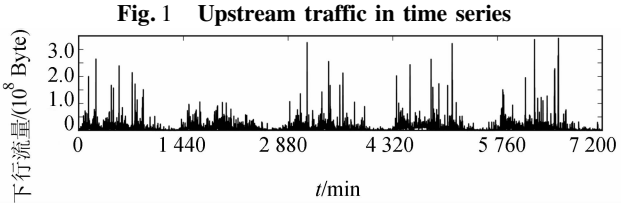

图 2　下行流量时间序列

Fig. 2　Downstream traffic in time series

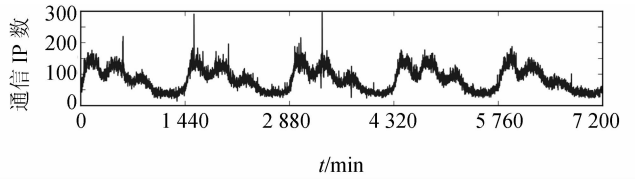

图 3 通信 IP 数量时间序列

Fig. 3 Communication IP count in time series

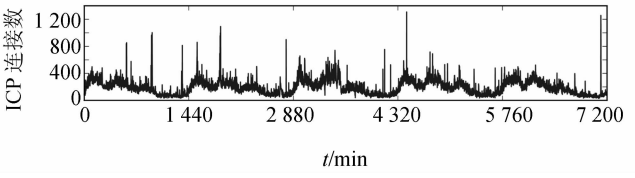

图 4 TCP 连接数时间序列

Fig. 4 TCP connection count in time series

从图 1 ~ 4 中可以看出,通信 IP 个数存在明显的周期性,其他的流量特征也存在周期性,但由于异常值得存在,导致周期性不明显,剔除这些异常值之后,整个服务器的流量将会非常稳定而有周期性。当网络中出现攻击、扫描、探测等异常行为时,网络中某些流量属性会发生变化。本文选取了表 1 中的 17 个特征,如非特别指明,表中特征都需要区分主动连接、被动连接。

表 1 特征列表

Tab. 1 Feature table

编号	特征	描述
1	TCP_Count	TCP 会话建立数
2	Send_Pkt_Count	上行数据包数量
3	Recv_Pkt_Count	下行数据包数量
4	Send_Len	上行流量
5	Recv_Len	下行流量
6	IP_Count	通信 IP 数量
7	Send_SYN_Count	发送 SYN 包数
8	Recv_SYN_Count	接收 SYN 包数
9	Send_FIN_Count	发送 FIN 包数量
10	Recv_FIN_Count	接收 FIN 包数量
11	Send_RST_Count	发送 RST 包数量
12	Conn_Per_IP	平均每个 IP 拥有的连接数
13	Send_Pkt_Len_Mean	平均发包长度
14	Recv_Pkt_Len_Mean	平均收包长度
15	TCP_UDP_Rate	TCP/UDP 比例
16	TCP_Fail_Count	TCP 会话建立失败数
17	Server_Port_Count	被动连接服务器端口访问数

如表 1 所示,流量特征根据 TCP 连接方向区分主动连接及被动连接,其中 TCP 会话建立失败是 TCP 3 次握手未成功的次数。

2 时间序列特性分析

通过第 1 节的特征选择,就可以得到多维特征随时间变化序列,以表 1 中第 1 号特征 TCP 会话数为例,描述对特征随时间变化序列进行时间序列分析的过程,该过程主要包括时间序列预处理、时间序列分析两个部分。

2.1 时间序列预处理

网络流量训练样本中往往存在异常数据,这往往是由于采集端错误造成的,这部分异常数据会影响时间序列分析算法的参数选择。根据格拉布斯准则[13],对网络流量数据进行适当预处理,平滑掉异常数据。

格拉布斯准则是判断数据粗大误差的一个准则,是异常数据剔除的常用方法。这里的具体做法是,把相同工作日对应的相同时刻的数据表示为 X_i,其中,$i = 1,2,3,4$。假设从一个月中的数据提取出完整的 4 周数据,所以,相同工作日、相同时刻的数据有 4 个,则表示 X_1、X_2、X_3、X_4 的平均值。

$$\bar{x} = \frac{1}{4}\sum_{i=1}^{4} x_i \tag{1}$$

$$v = \sqrt{\frac{1}{4}\sum_{i=1}^{4}(x_i - \bar{x})^2} \tag{2}$$

如果 X_i 满足:

$$|x_i - \bar{x}| > kv \text{ 且 } |x_i - \bar{x}| = \max_{j=1,2,3,4} |x_j - \bar{x}| \tag{3}$$

则认为 x_i 是坏值,应该剔除,其中,k 为格拉布斯准则系数,与 95% 置信区间相对应的 $k = 2.03$。

在实际计算过程中,如果训练样本数量较少,容易出现标准差极大的情况,许多显著异常值无法过滤。所以本文在预处理的过程中,采用了以下算法:

1)如果标准差大于 2 倍均值时,则剔除距离均值最大的值,重新计算均值、标准差。

2)每次循环只剔除一个坏值,下一次重新计算均值、标准差,直到所有的值都满足式(3)。

2.2 时间序列分析

在目前网络流量预测中,比较常见的时间序列模型有自回归模型(auto regression,AR)、滑动平均模型(moving average,MA)和两者的结合体自回归滑动平均模型(ARMA)。ARMA 适用于平稳时间序列的预测,而一般的时间序列往往不平稳,实际应用中常常采用差分操作,将非平稳时间序列转换成平稳型时间序列,所以此时的时间序列模型就是差分整合移动平均自回归模型(autoregressive integrated moving average model,ARIMA)。文献[14]采用季节 ARMA 模型

对通信网络中的异常点进行检测,通过预测值动态确定阈值,本文更倾向于短期预测,并不需要季节相关信息,因此采用 ARIMA 模型。

对于特征集中每一个特征随时间变化序列,根据其时间序列图、自相关函数和偏自相关函数识别其平稳性,对于非平稳时间序列,进行差分等平稳化处理,然后拟合 ARIMA 模型,确定 ARIMA 模型参数值,最后通过 ARIMA 模型进行预测。

对于 AR 模型,存在以下关系:

$$X_t = c + \sum_{i=1}^{p} \varphi_i X_{t-i} + \varepsilon_t \tag{4}$$

其中,c 为常数项,φ_i 为参数项,ε_t 为白噪声扰动项。AR 模型代表当前时刻值 X_t 与历史值 $X_{t-1}, X_{t-2}, \cdots, X_{t-p}$ 存在相关关系。如果当前值与以前时刻的扰动 ε_t 存在相关关系,则时间序列可以使用 MA 模型表示:

$$X_t = \mu + \varepsilon_t + \sum_{i=1}^{q} \theta_i \varepsilon_{t-i} \tag{5}$$

其中,X_t 为相关随机变量,θ_i 为参数项,ε_t 为白噪声扰动项。ARMA 模型则是上述 2 个模型的结合,最终 ARMA 模型可以表述为:

$$X_t = c + \varepsilon_t + \sum_{i=1}^{p} \varphi_i X_{t-i} + \sum_{i=1}^{q} \theta_i \varepsilon_{t-i} \tag{6}$$

ARMA 模型表示系统在 t 时刻的值 X_t 不仅与历史值 $X_{t-1}, X_{t-2}, \cdots, X_{t-q}$ 存在关系,还与随机扰动存在一定的关系。

根据 Box-Jenkins 提出的模型识别方法[15],平稳化处理后,如果偏自相关函数滞后 p 阶后截尾,而自相关函数拖尾,则建立 AR 模型;如果偏自相关函数拖尾,而自相关函数滞后 q 阶后截尾,则建立 MA 模型;若偏自相关函数和自相关函数均是拖尾的,时间序列适合建立 ARMA 模型。

计算结果如图 5 所示,从图 5(a)中可以看出,时间序列在滞后 1 阶之后不再显著,而从图 5(b)中可以看出,时间序列在滞后 2 阶之后不再显著,证明该时间序列平稳,可以使用 ARIMA 模型进行预测,而且 ARIMA 参数 p 取值为 1,q 取值为 2,d 表示差分次数,这里取值为 1。

图 6 是采用 ARIMA(1,1,2)对被动连接 TCP 会话数序列进行 120 次一步预测的结果。从图 6 中可以明显看出,在 32 min 的时候存在一个明显的异常点,真实值和预测值的偏离度较大。

通过对特征时间序列进行分析,证明特征时间序列在经过差分操作后,可以变成稳定型时间序列,

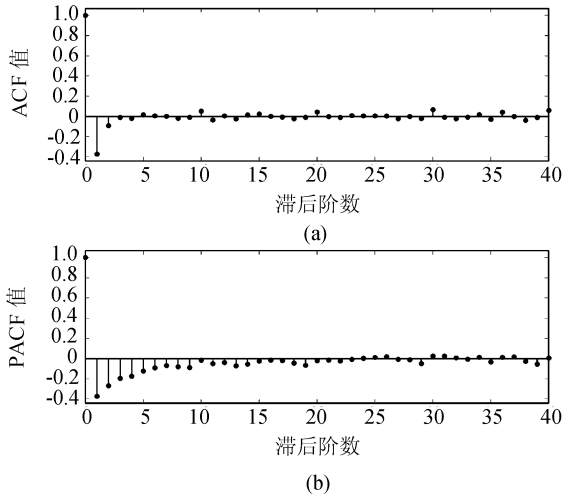

图 5　被动连接 TCP 会话数序列 ACF、PACF

Fig. 5　ACF and PACF series of passive TCP session

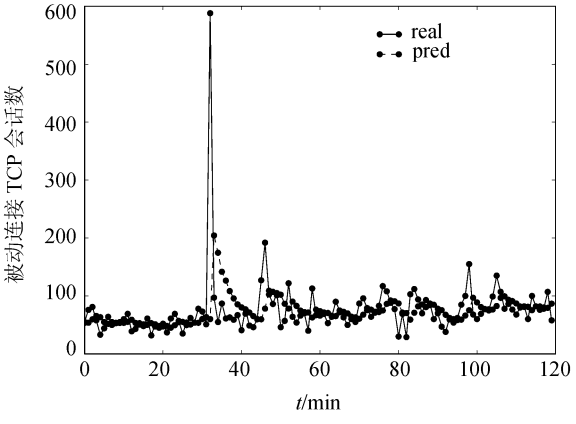

图 6　被动连接 TCP 会话数序列预测结果

Fig. 6　Prediction result of passive TCP series

从而利用 ARIMA 算法进行预测,由于预测值可以反映该特征时间序列的正常情况,当异常发生时,真实值将会偏离预测值。

3　异常检测方法

网络中的扫描、流量异常、主机探测等异常行为都会造成网络流量特征发生变化,根据第 2 节的时间序列分析,可以使用时间序列预测的方式衡量当前的正常情况,当真实值与预测值的偏离程度较大时,则认为网络中发生了异常事件。

3.1　时间序列偏离度

残差是真实值和预测值的偏差,它会在零均值附近波动,如果直接使用残差作为偏离度会存在以下问题:异常与残差的变动程度相关,各个特征的波动情况不一样。有些特征的残差虽然很高,但是总体波动幅度较大,如果单纯根据阈值判断,会造成误判。因此,本文使用时间序列偏度来度量当前网络流量与正常流量的偏离度,对每个特征的残差序列

进行分析,利用残差序列的整体情况来衡量当前残差,并定期更新残差序列。

定义 3(时间序列偏离度) 表示特征属性当前残差在整体残差序列中的偏离情况,对于残差序列 $e_1,e_2,\cdots,e_t,t=1,2,\cdots,N$,时间序列异常偏离度的计算公式如下所示:

$$\delta_t = \exp\frac{|e_t - \bar{e}_{t-1}|}{\sigma_{t-1}} \tag{7}$$

式中,e_t 为 t 时刻的残差,$\bar{e}_{t-1}$ 为 $t-1$ 时刻前残差序列的均值,σ_{t-1} 为 $t-1$ 时刻前残差序列的标准差。当残差与残差序列均值偏差越大时,表示真实值与预测值的偏差越大,采用指数运算更加放大了这种偏差。而且还统一了各个特征的量纲。

定义 4(时间序列偏离度向量) 对于多维特征,所有特征在时间段 t 内的偏离度构成时间序列偏离度向量,偏离度向量可以反映当前网络中流量与历史流量的偏离情况。

3.2 时间序列偏离度更新

当系统检测到异常值时,为了不让异常残差值对以后的偏离度计算产生干扰,需要对异常值进行处理。文献[3]使用残差序列中残差值的平均值替换异常值。虽然这样做可以减少异常值的影响,但是忽略了残差的趋势性。当真实网络环境中出现业务更新、新业务上线等情况时,网络中的流量也会随之发生变化,采用均值的方式就忽略了这种变化趋势。

基于上述原因,本文提出了残差序列更新算法,规定了在残差过大或者过小情况下的更新情况。每一次预测完成之后,都需要更新残差统计值。

1)残差 e_t 高于置信区间

$$e_t = \bar{e}_{t-1} + \alpha \times \sigma_{t-1},\ \text{if}\ e_t > \bar{e}_{t-1} + \alpha \times \sigma_{t-1} \tag{8}$$

式中:e_t 表示 t 时刻的残差;$\bar{e}_{t-1}$ 表示 $t-1$ 时刻前残差序列的均值;σ_{t-1} 表示 $t-1$ 时刻前残差序列的标准差;常数 α 表示对残差上界的容忍程度,其值越大,表明在更新残差序列时幅值越大。

2)残差 e_t 低于置信区间

$$e_t = \bar{e}_{t-1} - \beta \times \sigma_{t-1},\ \text{if}\ e_t > \bar{e}_{t-1} - \beta \times \sigma_{t-1} \tag{9}$$

式中:e_t 表示 t 时刻的残差;$\bar{e}_{t-1}$ 表示 $t-1$ 时刻前残差序列的均值;σ_{t-1} 表示 $t-1$ 时刻前残差序列的标准差;常数 β 表示对残差下界的容忍程度,其值越大,表明在更新残差序列时幅值越大。

3)残差在置信区间范围内

当残差在置信区间内或者经过第1)、2)步处理之后,对残差序列的当前的均值、标准差进行更新。

$$\bar{e}_t = \frac{\bar{e}_{t-1} \times (t-1) + e_t}{t} \tag{10}$$

$$\sigma_t = \sqrt{\frac{(t-1)(\sigma_{t-1}^2 + \bar{e}_{t-1}^2) + e_t^2}{t} - \bar{e}_t^2} \tag{11}$$

式(10)、(11)中 t 表示残差序列的长度,当 e_t 在置信区间外时,更新时的 e_t 需要经过式(8)或者式(9)处理。

3.3 支持向量机

支持向量机(support vector machine,SVM)算法在机器学习领域中被广泛使用,并取得了比较好的效果。它的基本思想是正确区分数据并且使分离超平面的几何间隔最大,最大间隔可以保证对未知的新实例有很好的分类预测能力。

支持向量机实质上是在约束条件下求解一个凸二次规划问题,通过拉格朗日对偶性变换到对偶变量的优化问题,通过求解与原始问题等价的对偶问题得到原始问题的最优解。对于每一个不等式约束,引进拉格朗日乘子 α,定义拉格朗日函数:

$$L(w,b,\alpha) = \frac{1}{2}\|w\|^2 - \sum_{i=1}^{N}\alpha_i y_i(w \cdot x_i + b) + \sum_{i=1}^{N}\alpha_i \tag{12}$$

根据拉格朗日对偶性,原始问题的对偶问题是极大极小问题:

$$\max_{\alpha}\min_{w,b} L(w,b,\alpha) \tag{13}$$

而对于非线性可分的情况,通过使用核函数变换以及引入松弛变量,最终可以在高维空间变得线性可分,从而使用线性可分方法进行分类。

在本文算法中,利用SVM算法对偏离度序列进行分类决策,该算法应用广泛,分类效果较好,而且支持多类分类。

3.4 异常检测过程

通过第2节的分析,使用ARIMA算法可以对流量特征时间序列进行预测,当网络中发生异常时,一个或者多个特征属性会发生变化,通过对多个属性偏离度,可以区分出不同的异常,从而检测出异常事件。

图7是本文所使用的多维特征向量,其中,$\boldsymbol{D}_{11},\boldsymbol{D}_{21},\cdots,\boldsymbol{D}_{nn}$ 代表偏离度序列,$f_1,f_2,\cdots,f_n$ 代表特征序列,而 $T_1,T_2,\cdots,T_n$ 代表时间序列,对于一个特征 f,都有其自己的ARIMA模型,都需要使用模型识别、模型定阶、模型参数估计,最终确定ARIMA参数。

本文使用多维特征时间序列分析进行异常检测关键步骤如下:

	T_1	T_2	T_3	…	T_n	
f_1	$\boldsymbol{D}_{11}$	$\boldsymbol{D}_{12}$	$\boldsymbol{D}_{13}$	…	$\boldsymbol{D}_{1n}$	ARIMA(p_1,d_1,q_1)
f_2	$\boldsymbol{D}_{21}$	$\boldsymbol{D}_{22}$	$\boldsymbol{D}_{23}$	…	$\boldsymbol{D}_{2n}$	ARIMA(p_2,d_2,q_2)
f_3	$\boldsymbol{D}_{31}$	$\boldsymbol{D}_{32}$	$\boldsymbol{D}_{33}$	…	$\boldsymbol{D}_{3n}$	ARIMA(p_3,d_3,q_3)
⋮	⋮	⋮	⋮		⋮	⋮
f_m	$\boldsymbol{D}_{m1}$	$\boldsymbol{D}_{m2}$	$\boldsymbol{D}_{m3}$	…	$\boldsymbol{D}_{mn}$	ARIMA(p_m,d_m,q_m)

图 7　多维特征向量

Fig. 7　Multiple feature vectors

Step 1:对于某一流量特征 x,首先确定 ARIMA 参数,然后利用 ARIMA 模型根据历史数据 x_{t-1},x_{t-2},…,x_{t-p} 预测 t 时段内流量特征预测值 y_t。

Step 2:利用式(7),计算 t 时段真实值 x_t 和预测值 y_t 的时间序列偏离度,当出现异常时,使用式(8)、(9) 调整历史残差序列,同时更新残差序列的标准差、均值等统计值。

Step 3:对于特征集合 $\{f_1,f_2,f_3,\cdots,f_n\}$ 中每一个特征,重复 Step 1、Step 2 的操作,直到得到各个特征在时间段 t 内的偏离度,构成偏离度向量 $\boldsymbol{D}_{1t}$,$\boldsymbol{D}_{2t}$,…,$\boldsymbol{D}_{mt}$。

Step 4:将 t 时段内的偏离度向量输入分类器进行分类判别,分类器根据各个特征维度上的偏离情况,决定异常类别。

4　实验结果及分析

4.1　实验数据

为验证本文算法的有效性,对四川大学某服务器流量进行异常检测。通过交换机端口镜像的方式,获得服务器区的原始流量,数据采集时间为 2016 年 4 月 25 日到 2016 年 4 月 29 日总共 5 d 时间,统计间隔为 1 min,总共 7 200 个时间段。流量总大小约 3.2 TB,其中,TCP 流量约为 2.6 TB,TCP 会话流数量约为 8 464 万条。结合网络安全设备日志,并且通过手工和机器结合的方式对这些时间段内服务器流量进行了异常识别、标注,结果如表 2 所示。

表 2　样本中各类异常数量

Tab. 2　Count of anomaly in sample

异常类别	异常数量
端口扫描	11
暴力破解	12
频繁连接	18
流量异常	38

表 2 中,针对相邻时间段内同一类型的异常,计作多次,并不合并为一次异常。例如,某次攻击持续 3 min,那么,计为 3 次异常。其中流量异常是某些时刻流量属性激增或骤减,但并不是网络攻击造成的。频繁连接是指网络中出现大量的 TCP 连接,但是异常规模又不构成 Dos 攻击,通常这些都属于主机探测。

在真实的校园网络环境中,几乎每天都有针对服务器的各种探测、密码破解等攻击行为,利用手工分析原始流量和安全设备检测日志等方式,对样本时间段内的异常进行识别,可以确保以识别出了所有的异常流量。

4.2　相关参数设置与确定

根据 Box-Jenkins 的模型识别方法,ARIMA 的参数可以使用自相关函数和偏自相关函数的拖尾特性进行判断,然后使用 Python 提供的 ARIMA 库对各时间序列进行参数估计,文中各特征的 ARIMA 模型参数选取结果如表 3 所示。

表 3　AMIRA 参数设置表

Tab. 3　ARIMA model parameter fit

编号	特征	p	d	q
1	单位时间 TCP 会话数	7	1	3
2	单位时间上行数据包数量	1	1	4
3	单位时间下行数据包数量	1	1	4
4	单位时间上行流量	1	1	4
5	单位时间下行流量	1	1	4
6	单位时间通信 IP 数量	1	1	4
7	单位时间接收 SYN 包数量	1	1	4
8	单位时间发送 FIN 包数量	1	1	4
9	单位时间接收 FIN 包数量	1	1	4
10	单位时间发送 RST 包数量	7	1	7
11	平均每个 IP 拥有的连接数	1	1	3
12	被动连接服务器端口访问数	1	1	1
13	单位时间 TCP 会话建立失败数	2	1	2

表 1 中,第 13 ~ 15 号特征值在训练样本中几乎不随时间发生变化,趋于固定值,因此采用固定阈值的方式对其检测,超过阈值标记为 1,其他标记为 0,然后输入 SVM 分类器。

4.3　检测结果分析

为了验证本文提出的时间序列偏离度算法的准确性,与文献[4]在构建异常向量的过程中采用的 EWMA(exponentially weighted moving average)预测算法进行对比。使用 EWMA 算法计算各个维度上时间序列的预测值与实际值的偏差构成异常向量,与本文使用的残差偏离度向量同时输入到 SVM 中进行异常检测,实验结果如表 4 所示。

表4 算法实验效果对比表
Tab.4 Comparison of algorithm

异常类别	本文方法		EWMA方法	
	检测率/%	误检率/%	检测率/%	误检率/%
端口扫描	90.9	0	81.8	9.1
暴力破解	91.6	8.3	83.3	8.3
频繁连接	88.2	16.7	72.2	11.1
流量异常	92.1	13.1	84.2	10.5

从表4中可以看出,本文算法检测率总体上优于EWMA算法,尤其是针对流量异常和暴力破解。主要是因为:1)本文使用了ARIMA算法,该算法预测精度相比于EWMA算法更高,虽然在模型训练阶段需要较大的计算量,但是预测阶段计算量较小。2)针对不同的特征属性进行预测,然后将各个时间段内特征组合成检测向量输入到SVM分类器进行分类,能够有效提高检测率。但是针对频繁连接和流量异常,本文算法仍具有一定的误报率,这主要是由于标注过程中,部分频繁连接和流量异常特征上相似造成的。

5 结 语

通过对真实网络流量的长期观测,提出了随时间变化的多维特征序列,通过对历史特征值序列进行分析,得到了反映当前网络情况的预测值,计算预测值与真实值的偏离度,得到多维特征偏离度向量,最后使用训练的SVM算法对偏离度向量进行分类、判别。在真实的网络环境中进行实验,结果表明该方法可以有效地检测网络中的异常流量。

在下一步的工作中,将异常检测算法进行扩展,不只是研究单台主机网络行为,而是关注于整个网络以检测全局网络中的异常流量。

参考文献:

[1] Roy D B, Chaki R. State of the art analysis of network traffic anomaly detection[C]//Applications and Innovations in Mobile Computing (AIMoC). Kolkata: IEEE, 2014: 186 – 192.

[2] Bhuyan M H, Bhattacharyya D K, Kalita J K. Network anomaly detection: Methods, systems and tools[J]. Communications Surveys & Tutorials, IEEE, 2014, 16(1): 303 – 336.

[3] Lv Junhui, Zhou Gang, Jin Yi. Adaptive aberrant network traffic detection algorithm based on time series forecast[J]. Journal of Beijing University of Aeronautics and Astronautics, 2009, 35(5): 636 – 639. [吕军晖, 周刚, 金毅. 一种基于时间序列的自适应网络异常检测算法[J]. 北京航空航天大学学报, 2009, 35(5): 636 – 639.]

[4] Ye Xiaoming, Chen Xingshu, Wang Haizhou, et al. An anomalous behavior detection model in cloud computing[J]. Tsinghua Science and Technology, 2016, 21(3): 322 – 332.

[5] Zheng Liming, Zou Peng, Han Weihong, et al. Traffic anomaly detection using multi-dimensional entropy classification in backbone network[J]. Journal of Computer Research and Development, 2012, 49(9): 1972 – 1981. [郑黎明, 邹鹏, 韩伟红, 等. 基于多维熵值分类的骨干网流量异常检测研究[J]. 计算机研究与发展, 2012, 49(9): 1972 – 1981.]

[6] Wang Shuo, Zhao Rongcai, Shan Zheng. Distributed denial of service detection algorithm based on FSS time series analysis [J]. Computer Engineering, 2012, 38(12): 13 – 16. [王硕, 赵荣彩, 单征. 基于FSS时间序列分析的DDoS检测算法[J]. 计算机工程, 2012, 38(12): 13 – 16.]

[7] Bai Jun, Xia Jingbo, Zhang Wenjing, et al. Rapid botnet detecting method based on multi-dimensional information divergence[J]. Journal of Huazhong University of Science and Technology (Natural Science Edition), 2014, 42(9): 28 – 32. [柏骏, 夏靖波, 张文静, 等. 基于多维信息散度的僵尸网络快速检测方法[J]. 华中科技大学学报(自然科学版), 2014, 42(9): 28 – 32.]

[8] Lai Yingxu, Jiaojiao. Anomaly detection scheme using time sereis analysis for industrial control systems[J]. Journal of Beijing University of Technolog, 2015, 41(2): 200 – 206. [赖英旭, 焦娇. 基于时间序列分析的工业控制以太网流量异常检测[J]. 北京工业大学学报, 2015, 41(2): 200 – 206.]

[9] Shirani P, Azgomi M A, Alrabaee S. A method for intrusion detection in web services based on time series[C]//2015 IEEE 28th Canadian Conference on Electrical and Computer Engineering (CCECE). Halifax: IEEE, 2015: 836 – 841.

[10] Wu Na, Mu Zhaoyang, Zhang Liangchun. Distributed denial of service convert flow detection based on data stream potential energy feature[J]. Computer Engineering, 2015, 41(3): 142 – 146. [吴娜, 穆朝阳, 张良春. 基于数据流势能特征的分布式拒绝服务隐蔽流量检测[J]. 计算机工程, 2015, 41(3): 142 – 146.]

[11] Moore A, Zuev D, Crogan M. Discriminators for use in flow-based classification[R]. Cambridge: Queen Mary University of London, 2005.

[12] Li Xiulong. Research on user's traffic behavior analysis method based on network traffic monitoring and prediction [D]. Beijing: Beijing University of Technology, 2013. [李秀龙. 基于网络流量监测与预测的用户流量行为分析方法研究[D]. 北京: 北京工业大学, 2013.]

[13] 费业泰. 误差理论与数据处理[M]. 北京: 机械工业出版社, 2000.

[14] Yu Yanhua, Song Meina, Zhang Wenting, et al. A dynamic computation approach to determining the threshold in network anomaly detection[J]. Journal of Beijing University of Posts and Telecommunications, 2011, 34(2): 45 – 49. [于艳华, 宋美娜, 张文婷, 等. 网络异常点检测中性能指标阈值的动态确定方法[J]. 北京邮电大学学报, 2011, 34(2): 45 – 49.]

[15] Cryer J D, Chan K S, 潘红宇. 时间序列分析及应用: R语言[M]. 北京: 机械工业出版社, 2011.

(编辑 杨 蓓)

引用格式: Chen Xingshu, Jiang Tianyu, Zeng Xuemei, et al. Network anomaly detector based on multiple time series analysis [J]. Advanced Engineering Sciences, 2017, 49(1): 144 – 150. [陈兴蜀, 江天宇, 曾学梅, 等. 基于多维时间序列分析的网络异常检测[J]. 工程科学与技术, 2017, 49(1): 144 – 150.]