基于流记录的异常网络流量检测

吴琪 丁伟

（东南大学 计算机科学与工程学院， 南京 211189）

摘要 异常网络流量行为的检测对于提高网络的可用性和可靠性有着重要的意义。因此，这是一个重要的研究领域，持续多年的研究发展出很多经典的网络异常检测算法。本文基于一个精细化的网管平台 NBOS 提供的细粒度的分类流量信息，对现有的相关算法进行了分析，确定 Holt-winters 模型最适合于 NBOS 平台环境。将该算法集成进 NBOS 平台并部署在 CERNET 南京主节点后，在实际的流量环境中实时检测出了大量异常流量案例，论文对其进行了展示和分析。

关键词 异常检测；流记录；时间序列模型；指数平滑； Holt-winters 模型

Network Traffic Anomaly Detection Based on Flow Records

Wu Qi DING Wei

(School of Computer Science and Engineering, Southeast University, Nanjing 211189)

Abstract Network traffic anomaly detection is much important to improve network availability and reliability. Hence,it is a topic of concern and many classical anomaly detection algorithms have been developed.Based on the fine-grained classified traffic data provided by NBOS,which is a fine-grained network management system, the paper analysed existed related algorithms and determine that method based on Holt-winters is most appropriate for the platform environment of NBOS.The method is deployed at NanKing node of CERNET and a massive of network traffic anomalies cased have been detected . Relevant demonstration and analysis will be displayed in the paper.

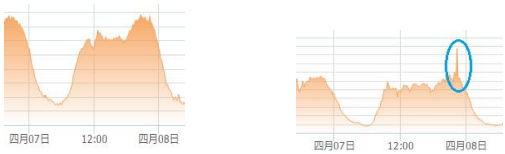
Keywords anomaly detection ; flow records; time sequence model; exponential smoothing; Holt-winters model

1 引言

大量研究表明,网络流量具有自相似、长相关和重尾分布等分布特征[1]。这些研究结果对网络流量工程、网络建模和异常检测具有指导意义。网络流量在正常运行的情况下是具有一定的周期性和稳定性的，当流量偏离了其正常行为，打破了这种规律时，就可能发生了异常。引发主干网流量异常的原因可分为三类[2]：网络攻击异常，如 DDoS、扫描等；网络设备故障异常，如路由器故障导致流量过载、网络拥塞；瞬间大量访问异常，如节假日的网页访问量猛增。无论哪一种原因，准确、快速地检测到这些行为并做出合理的响应,可以保障网络的安全运行，因此这是一个有意义的研究领域，持续多年的研究出现了很多经典的检测算法[3-6]。

然而，这些经典的异常流量行为检测算法却很少真正进入工程领域。可能的原因之一是目前流量分析的数据源相对单一，基本来自网络设备提供的 SNMP 数据。SNMP 是一种广为使用的网络协议，它使用嵌入到网络设备中的代理软件收集网络通信信息和有关网络设备的统计数据并将其记录到管理信息库（MIB）中。这些信息中与带宽有关的部分包括占用带宽和字节数，它们是根据链路层地址进行聚合的，无法反映报文中 IP 地址和端口号等信息。在这样的数据条件下使用异常流量行为算法仅能对网络设备端口的整体宏观流量进行分析,可能会漏报一些局部的异常。图 1 是 NBOS 系统提供的同一时间段的流量信息，(a)是 CERNET 江苏省网的整体流量情况，这个带宽数据 SNMP 也可以提供；(b)是省内某 985 大学的流量，这个带宽数据是 CERNET 江苏省网边界路由器无法提供的。需要强调的是(b)中尖峰点在(a)中没有体现，这是因为异常体现的是一个相对的关系。同时也表明 NBOS 对带宽的细粒度的刻画是有分析价值的。

NBOS 系统可以对带宽进行细粒度的刻画的原因是其使用路由器的流记录作为



(a)整体流量　　(b)省内某 985 高校流量

图 1 NBOS 系统提供的流量信息

分析数据源。NBOS [7] (Network Behavior Observation System)是在 CERNET 211 三期工程支持下开发的一个面向流记录的综合处理平台。流是指在特定时限内具有共同特征的一组数据包[8]。与 SNMP 数据相比，流记录数据能提供更详细的网络层及其以上的协议数据，包括协议类型、AS 号、协议端口号等。虽然在实际运行环境中，流记录通常基于抽样报文产生，但相关研究已经证明,抽样不会影响异常数据包在流量中的比例,因而不会对异常检测产生重大影响[9]。NBOS 能够基于主干链路上被动测量获取的流数据，以不同的时空尺度提供网络的基础运行数据，为网络的异常流量检测提供比 SNMP 更理想的分析数据源。

本文将首先针对异常网络流量行为和相关的检测算法展开研究。在此基础上，基于 NBOS 可以提供的基本流量行为数据，验证经典异常检测算法在大规模高速网络环境中面向不同时空尺度的分析数据源的实际运行效果。

2 NBOS 的带宽测度

与 SNMP 不同的是，NBOS 提供的带宽测度通过从网络边界（以下简称为接入点）的角度对 IP 地址空间进行划分后，以更精细的粒度提供。接入点以内的网内 IP 地址空间按其所归属的接入单位(比如，XX 大学)划分成分析对象，整个接入网（以下简称全网），也被看成是一个分析对象；接入点以外的 IP 地址空间按教育网地址、国内运营商地址、非中国大陆地址和其他地址划分成四类。这样，NBOS 提供的带宽数据可用 Object_Bandwidth =U:F 进行描述, U 是分析对象，就是上文所述的接入单位；F 是测度，它由一个属性三元组(A,B,C)构成，A={CERNET,大陆运营商，非大陆地区，其他，总体}，称为对端属性，其中总体是其他四类数据之和，为了方便表述，分别用 1、2、3、4、5 表示，B={入方向，出方向}，称为方向属性,分别用 i 和 o 表示，C={bps,pps}，称为统计单位属性,分别用 b、p 表示。因此对于每个对象 U（比如，XX 大学），NBOS 可以提供的带宽测度总数为 |A|*|B|*|C|=5*2*2=20 个，每个对象的每个测度都可以用一张以时间为横轴的带宽折线图展示，如图 1 中(a)的展示对象 U=Net，代表全网，(b)的展示对象 U=X，代表 XX 大学，展示的测度均是(总体，出方向，pps)，分别表示为 Net:5op 和 X:5op。在本文后面的部分，我们将多次使用这样的表述。

对于整个接入网或任何一个接入单位，NBOS 为异常检测提供了比 SNMP 更为精细的分析数据源。据此，本文的研究目标是基于上述数据，尝试对不同的测度选取最适当异常流量行为检测算法，并将其集成 NBOS 系统，用于在更加精细地检测网络流量中的异常流量行为。

3 检测算法的分析

3.1 流量异常检测方法

异常检测依赖于“异常”的定义。Maxion R.A 给出的关于网络的“正常”和“异常”描述如下[10]：“正常”意味着符合某种常规或典型的模型，以一种自然的方式，常规的或预料中的状态、形式、数量或程度发生，“正常”强调符合某种已经建立的水准或模式，并保持良好状态，建立在一定趋势基础上。而“异常”意味着违反了这种期望，与期望的情形有一定程度的偏差。

异常检测首先建立网络流量正常状态的基准轮廓。然后，将网络流量的当前状态与基准轮廓进行比较，偏离正常基线的状态将被认为是异常。通用的异常检测系统框架如图 2 所示。

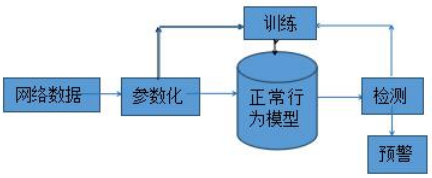


图 2　异常流量检测模型

根据采用的检测方法，异常检测算法可分为基于统计分析、基于小波分析、基于机器学习、基于数据挖掘等。基于统计分析的方法是按一定的时间间隔对流量行为进行采样，通过计算采集样本得到一组参数变量，从而产生表示流量行为的轮廓，将每次采样后得到的行为轮廓与已有轮廓进行合并，最终得到正常的行为轮廓。通过将当前

采集到的行为轮廓与正常行为轮廓相比较，来检测是否存在异常行为。Denning[11]提出了用于异常检测的 5 种统计模型：操作模型；方差；多元模型；马尔可夫过程模型；时间序列模型。统计检测方法在实际应用中有多种算法[5，6]。

3.2 面向分析数据特征的分析算法分类

分析算法的分类可以从不同角度进行。根据所处理的数据的特征，可将分析算法可以分为两类[3,13,14]。

a.基于流量量值

基于流量量值的异常检测，统计流量在包数量、流量比特等方面的数据。文献[13]提出尽管网络的流量行为是非常不确定的, 但是通过观察可以发现, 每个流量参数在一周各天的观测值都呈现出周期性趋势, 这在该大型业务网络中也得到验证[15]。基于流量量值的异常检测方法得到广泛研究，它常常借助时间序列分析和信号处理等理论知识进行分析。NBOS 提供的 20 个测度属于这一类。

b.基于流量信息结构

不是所有网络异常流量行为都能够引起明显的流量量值的偏移。基于流量信息结构的异常检测技术，通过检测网络流量属性分布特征的变化检测异常。[13]中发现网络流量在 IP 地址和端口的分布上存在较强的重尾分布和自相似特性。网络流量在宏观上量值变化较大，但在微观层面具有较为稳定的分布结构。这类异常检测通常在流量基本特征的基础上，采用熵、神经网络等方法进一步提炼流量特征规律，在此基础上进行检测。除了上述 20 个测度外，NBOS 还能够提供各协议端口单位时间内报文数、字节数等流量信息，这些数据属于这一类。

3.3 分析算法的选择

研究工作以每个分析对象的 20 种测度为单位进行，基于 NBOS 平台实现在线检测。因为是研究工作的起步同时考虑实时性方面的要求，我们首先考虑计算量小、延迟时间短、误检率低的检测方法。

3.3.1 可使用的检测方法

小波分析作为在信号处理领域广泛使用的方法，虽然对异常网络流量行为有较好的检测效果，但计算过程复杂，有相当于滑动窗口大小的延迟，因此不适合在线检测。作为分析对象的 20 种测度是单位时间内的报文数和字节数，按一定的时间间隔进行采样，是一种典型的时间序列数据。基于时间序列分析的方法可以较好地揭示网络流量内在统计特征和变化规律，可以应用相关理论与模型来建立网络流量模型。时间序列分析理论认为这些观测值是不独立的，有一定的相关性，未来的数值可以由历史数据来预测。因此，可以通过这种相关性来建立相应的数学模型来描述网络流量的动态特征。具体方法包括移动平均法、自适应过滤法、状态空间法、时间序列分解法、指数平滑法等。

依据时间序列建立起来很多模型，常用的模型有 ARMA 模型、ARIMA 模型和指数平滑模型。AR 模型和 MA 模型都是 ARMA 模型的特例。GLR(Generalized Likelihood Ratio)就是在 AR 模型上建立的异常检测方法。它考虑 2 个相邻的时间窗 R (t) 和 S (t) 以及二者合并组成的窗口 C (t), 每个窗口都采用自回归 AR (Auto-Reg ressive)模型拟合, 计算各窗口序列残差的联合似然比, 当该值超过预先设定的阈值 T 时, 两个窗口 R (t)和 S (t)的边界就被认定为异常点。应用 ARIMA 模型进行异常检测也已有成熟的研究。这些方法实质上是用可以精确描述历史数据的流量模型来近似代替当前的流量模型。随着时间的推移与网络流量的随机波动，准确性会很快下降。在确立流量模型时，如何减少对历史数据的依赖性、适应流量的实时变化，同时增加检测算法的准确性，是一个关键问题。此外，这类方法计算量较大。因此我们选择计算开销小并且有较好自适应的基于指数平滑的异常检测算法进一步分析。

3.3.2 算法 1：基于指数平滑的网络异常检测

指数平滑法是 1959 年由美国学者布朗在《库存管理的统计预测》一书中提出来的，它不需要除序列自身以外的、其他序列的额外信息，是根据自身预测自身的一种预测方法。它利用历史数据和相关统计信息，根据厚近薄远的原则进行加权并修匀数据，因此该数据模型具有检索异常数据影响的功能，并能显著的体现出时间序列包含的历史规律。

利用指数平滑技术检测网络异常时，选取的网络参数通常为单位时间内的数据包的比特数、包数等[5]。记顺序到达的网络流量观测序列为 $x_1, x_2, \cdots x_{t-1}, x_t, x_{t+1} \cdots$。记 t-1 时刻该序列的预测值为 $\hat{y}_{t-1}$，实际观测值为 y_{t-1}，则 t 时刻的预测值为：

$$\hat{y}_t = \partial * y_{t-1} + (1-\partial) * \hat{y}_{t-1} \quad (1)$$

对其扩展，可得式(2)：

$$\hat{y}_t = \partial * y_{t-1} + (1-\partial) * [\partial * y_{t-2} + (1-\partial) * \hat{y}_{t-2}] \quad (2)$$

依次类推,可得式(3)：

$$\hat{y}_t = \partial * \sum_{i=2}^{t-1} (1-\partial)^{t-i-1} * y_i + (1-\partial)^{t-2} * \hat{y}_2 \quad (3)$$

权重 $\partial * (1-\partial)^{t-i-1}$ 随着 i 的变小呈几何级递减，所以较早数据对预测值的影响较小，离预测点较近的数据影响较大，这就是指数平滑得名的原因。

∂ 是平滑指数，决定了预测值对历史数据的衰减（快慢）程度。它是介于 0~1 之间的数，取值可以根据过去数据点所占的权重计算得出，方法如下：

$$\partial = 1 - \exp\left[\frac{\log(1-w\%)}{n}\right] \quad (4)$$

式中，log 表示自然对数；w 表示权重的百分比形式；n 表示所取时间序列数据点的数目。

α的取值对于预测至关重要。不同时期、不同序列对历史数据的依赖程度不同，α取值也应该不同。每一个时间序列，对其滑动窗口中的序列分别取α={0.15, 0.2, 0.25, 0.3, 0.35，0.45}求出预测序列，然后采用均方误差(MSE)极小的原则选取α值。MSE 公式见式(5)下，y_t 和 f_t 分别代表实际值和预测值，n 为滑动窗口中的序列。这个α值就是当前该序列最优的指数平滑系数。

$$MSE = \frac{1}{n}\sum_{t=1}^{n} (y_t - f_t)^2 \quad (5)$$

这种方法适用于序列值围绕自身均值上下作随机波动的序列，网络单位时间内的报文数和字节数在一定时间内往往围绕一个水平上下波动,因此可以通过该方法检测异常。

3.2.3 算法 2：基于 Holt-Winters 季节模型的网络异常检测

在统计学中，时间序列的变化分解为四种：趋势变化、周期变化、循环变化、随机变化。前两种有规律可循因而容易量化，后两者波动比较复杂而较难量化。时间序列的预测主要就是研究趋势变化和周期变化，并抑制循环变化和随机波动对模型建立的影响。在用算法 1 进行预测时，处理的时间序列中包含着以上四种变化，且这四种变化相互融合在一起，难以区分，往往影响到了检测效果。

文献[6]应用了Holt-winters线性季节平滑模型进行自适应的异常行为检测。文献[12]中提出“具有季节周期为 d 的 Holt-winters 递归表达与 ARIMA 过程的某个大样本预测递归相符”。同时与 ARIMA 相比，Holt-winters 模型更为简单，自适应能力快，更适合在线检测。

Holt-Winters 是基于指数平滑法的线性预测，它把预测值分为了 3 个部分：基线值 baseline、线性趋势 lineartrend 和季节影响 seasonal effect。t+1 时刻的流量预测值的计算公式为：

$$\hat{y}_{t+1} = a_t + b_t + c_{t-m+1} \quad (6)$$

其中，a_t，b_t 和 c_t 分别表示基线值、线性趋势和季节影响，d 表示季节周期。这 3 部分的更新方式如下：

$$a_t = \alpha * (y_t - c_{t-d}) + (1-\alpha)(b_{t-1} + a_{t-1}) \quad (7)$$

$$b_t = \beta(a_t - a_{t-1}) + (1-\beta)b_{t-1} \quad (8)$$

$$c_t = \gamma(y_t - a_t) + (1-\gamma)c_{t-d} \quad (9)$$

α、β、γ是算法的自适应参数。通过衡量预测值与实际值的偏离来检测数据是否发生异常。t 时刻的预测偏移 d_t 计算公式如下：

$$d_t = \gamma * |y_t - \hat{y}_t| + (1-\gamma) * d_{t-m} \quad (10)$$

y_t 的置信区间为($\hat{y}_t$ - δ_- * d_{t-m} ，$\hat{y}_t + \delta_+ * d_{t-m}$)。δ_- 和 δ_+ 是缩放因子，用来改变置信区间的大小。

在实际检测中，要选取合适的一组参数α、β、γ，才能达到良好的检测效果。由于模型中参数之间的相互作用，并没有绝对最优的一组参数，通常可以在 0.2 和 0.4 之间的几个值反复试验[6]，通过为实际数据和预测值的误差定义一个标准来选择一个合适的参数。α的含义同式（4），如果希望最近 45 分钟内的采样值(单位时间为 5 分钟，共 9 个采样点)占 95% 的权，那么α=1-exp(log(1-95%)/9) = 0.28；如果希望最近 1 小时内的观测值占 75%的权，那么α=1-exp(log(1-75%)/12) = 0.11。对α取 0.1 和 0.28 之间几个值实验比较效果。β的目标是获取超过季节影响的线性趋势，因而β应取较小的值。周期为 1 天，最近一天的采样值占权重为 50%时，β为 0.0024。γ控制季节性变化系数。在我们实际网络环境中测试，取α=0.28，β=0.0024,γ=0.1，能较好的为流量行为建模。

4 检测算法的选择

如图 2 所示，流量异常检测的核心是实现流量正常行为的描述。算法 1 和算法 2 都属于基于预测的异常检测，其效果直接依赖于建立的模型是否能够准确描述复杂多变的网络流量。本节将根据 NBOS 提供的实测数据，从中选择一个更加适合它的算法。

工作的思路是选择若干对象的特定测度，利用历史数据为其建立基准模型。将算法 1 和算法 2 同时作用于该模型，对它们建立的预测模型与基准模型进行比较。

4.1 对象和测度的选择

实验数据来自于 CERNET 南京主节点，该节点接入了 155 个单位。因此，根据第 2 节的内容，流量测度的分析对象一共有 156 个。我们从中选择两个代表性对象 A 和 B。其中，A 是某 985 高校，属于大流量单位，选择其 5ib；B 是某 211 高校，属于小流量单位，选择其 3op。

4.2 基准模型建立方法

我们需要一个正常流量行为的模型作为基准，在此基础上验证两种算法的预测效果。分析周期是“天”，而 NBOS 以 5 分钟为单位时间提供基本流量数据。这样对于每一个测度，每小时有 12 个采样点，每天有 24*12=288 个采样点，因此每个测度的基准数据有 288 个。

所有的指数平滑方法都是基于递推关系的，因而要设定初始值，对 Holt-Winters 而言，必须初始化一个完整的“季节周期”值。同时指数式衰减规律说明所有的指数平滑方法的“记忆”能力都很短，初始值的影响不大。本文设定算法 2 中 y_0 和 a_0 为检测开始时上一单位时间的采样值，b_0 和第一个季节周期的 c_{t-d} 全 0。因此，为了达到更好的验证效果，对算法 2 而言，至少需要 2 个周期的启动过程。对于算法 1，y_0 和 $\hat{y}_0$ 为检测开始时上一单位时间的采样值，只需要一个单位时间就可以启动。考虑到算法 2 的启动过程，基准模型的时间长度 T 设定为 3 天，前面 2 天的数据作为“学习”，对第 3 天的数据进行检测。

我们利用历史数据建立基准模型。考虑到历史数据中可能存在异常，我们对连续 7 天同一采样点的数据用下面的方法剔除坏值(其相对误差为粗大误差)后取均值作为基准。具体步骤如下：

Step1: 用 y_{ij} 表示选定测度过去 7 天的采样值，i 代表天，j 是采样点，i=1,2,...,7；j=1,2,...,288。

Step2: 对 j=1,2...,288，把全部 y_{ij} 数据(i=1,2..,7)去除最大值、次最大值、最小值，计算剩余 4 个数据的均值作为该采样点的基准值。

4.3 比较和选择

在实践中，我们用 2016 年 4 月 18 日 0 时至 26 日 24 时共 9 天的流量数据作为历史数据。对 18~24 日、19~25 日、20~26 日的数据通过 4.2 节中的方法分别计算出基准模型“第 1 天”、“第 2 天”、“第 3 天”的数据，即使用 9 天的历史数据计算出 3 天的基准模型。使用两种算法同时作用于该模型。运行到模型的“第 3 天”时，用式(11)统计两种算法288 个采样点单位时间预测值与实际值的偏差情况，y_j 为基准模型实际值，$\hat{y}_j$ 为算法的预测值，n 为序列值个数。

$$\zeta = \sum_{j=1}^{n} | y_j - \hat{y}_j | / n \qquad (11)$$

实验的具体方案如下：

Step1：获取 9 天历史数据。

Step2：剔除坏值、建立 3 天基准模型。

Step3：算法 1 和算法 2 同时作用该基准模型。

Step4：对模型“第 3 天”计算 ζ，选择 ζ 值较小的算法。

分析基准模型“第 1 天”选定测度的数据，统计情况如表 1 所示。ζ 的计算结果如表 2 所示。

表 1 流量数据统计信息

流量测度	最大值	最小值	均值
A:5ib	148571606016	201438720	7455270873 6
B:3op	3280896	66816	1011190

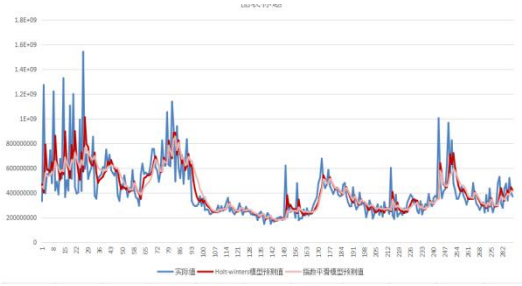
表 2 两种算法预测值与实际值的平均偏差

	算法 1	算法 2
A 选定测度	4839253837	4463830067
B 选定测度	126252	124508

两种算法“第 3 天”的运行情况如图 3 所示。其中，横坐标表示单位时间，蓝色、红色、粉色曲线分别表示采样值、算法 2 和算法 1 的预测值。

可以看出:不论作用于大流量测度 A:5ib 还是小流量测度 B:3op，算法 2 的 diff 值都

(a) A:5ib



(b) B:3op

图 3 两种算法的预测效果

较小。这说明基于 Holt-Winters 季节模型的检测方法能够更快速地适应流量变化、更准确地刻画流量的正常模型，因此本文选择算法 2 作为检测方法。

5 基于 NBOS 平台的实现

我们将算法 2 基于 NBOS 平台部署在 CERNET 南京主节点，在每个 NBOS 的采样点对 156 个分析对象的 20 个测度进行检测。我们对检测结果从三个方面进行分析。

5.1 24 小时结果统计

5 月 5 日 0 时至 24 时，实验的检测结果如表 3 所示。

表 3 5 月 5 日实验结果统计信息

异常测度	异常事件数	涉及分析对象数	异常测度	异常事件数	涉及分析对象数
1ip	1169	142	1op	1045	122
1ib	1193	143	1ob	1181	129
2ip	678	151	2op	426	119
2ib	707	148	2ob	468	116
3ip	684	140	3op	600	121
3ib	783	146	3ob	614	123
4ip	997	151	4op	1644	114
4ib	1088	149	4ob	1920	117
5ip	612	145	5op	557	119
5ib	683	142	5ob	605	123

5.2 准确性分析

本节选择一个分析对象，对检测出的异常测度进行验证。通过式(12)对这些发生异常的采样点及其之前采样点统计它们与相邻粒度的差异情况。

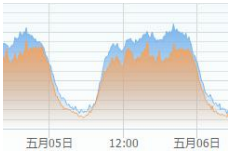
$$diff_y(t) = \left| \frac{y_t - y_{t-1}}{y_{t-1}} \right| \quad (12)$$

我们选择某 211 高校 C，5 月 5 日 0 时至 24 时 C 的检测结果如表 4 所示。

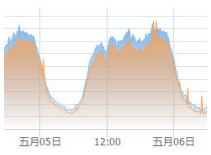
表 4 C 检测出的异常事件

序号	方向	异常测度	时间
1	出	3ob,5ob	01:30
2	入	4ib	03:45
3	入	4ib	06:45
4	入	1ip，1ib,5ip，5ib	11:00
5	出	1op,5op	11:00
6	出	2ob	14:40
7	出	3ob	14:55
8	出	1ob	15:15
9	入	4ib	16:15
10	入	4ib	22:55
11	出	3ob,5ob	22:55

图 4 是全网和 C 5 月 5 日的总体流量情况，蓝色曲线表示入方向，橙色曲线表示出方向。11 时左右，由(c)可知，C：5ib 发生明显的异常，C：5ob 无异常;由(d)可知，C：5ip 和 C：5op 均发生明显的异常；由(a)和(b)可以看到全网总体流量无异常。该分析对象总体流量的异常无法在全网流量中反映。对表 4 中的异常点计算 diff 值，结果如表 5 所示，采样点 1.1 表示表 4 中序号为 1 的一栏中第一个异常测度。可以发现：这些异常点的 $diff_y(t)$ 远大于该点之前的 $diff_y(t)$ 的平均值。



(a) 全网 bps 流量数据



(b) 全网 pps 流量数据

(c) C bps 流量数据

(d) C pps 流量数据

图 4 5 月 5 日全网和分析对象 C 流量数据

表 5　表 4 中异常点的 diff 值

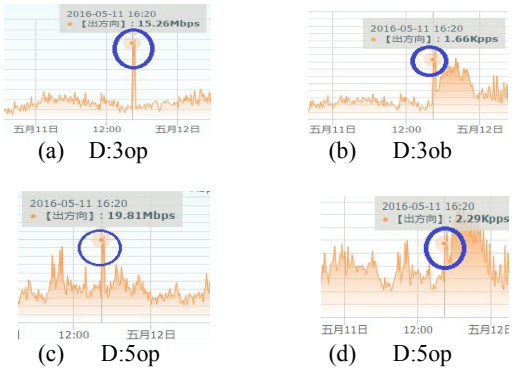
采样点	-4	-3	-2	-1	异常点
1.1	0.144	0.202	0.381	0.152	1.695
1.2	0.272	0.111	1.386	0.156	4.002
2	0.018	0.174	0.338	0.480	7.971
3	0.317	0.047	0.286	0.247	12.182
4.1	0.045	0.029	0.098	0.423	2.097
4.2	0.086	0.162	0.153	0.439	3.034
4.3	0.082	0.189	0.226	0.876	3.591
4.4	0.094	0.236	0.248	0.856	3.932
5.1	0.059	0.138	0.113	0.342	2.360
5.2	0.148	0.193	0.173	0.872	3.235
6	0.391	0.174	0.076	0.366	6.291
7	0.179	0.157	0.113	0.052	11.270
8	0.594	0.609	0.381	0.087	5.260
9	0.142	0.348	0.261	1.102	3.027
10	0.657	0.176	0.013	0.014	7.098
11.1	0.048	0.023	0.220	0.215	4.199
11.2	0.988	0.139	0.228	0.251	9.647

5. 3 案例分析

本节选择两个典型事例进一步分析。

5. 3. 1 案例一

5 月 11 日 16 时 20 分，分析对象 D 出方向上 4 个测度:3op,3ob,5op,5ob 检测到异常。这一段时间 D 流量数据如图 5 所示。可以发现：该时刻 D 出方向上非大陆地区流量发生了明显的异常,并反映在了 D 出方向总体流量上。



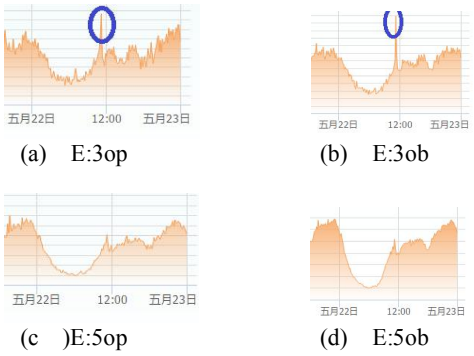
(a)　D:3op　(b)　D:3ob

(c)　D:5op　(d)　D:5op

图 5　D 5 月 11 日流量数据

5. 3. 2 案例二

5 月 22 日 11 时 10 分，分析对象 E 出方向上 2 个测度:3op,3ob 检测到异常。这一段时间 E 流量数据如图 6 所示。可以发现：该时刻 E 出方向上非大陆地区流量也发生了明显的异常，但不能在 E 出方向总体流量上反映。这说明 NBOS 为异常检测提供了理想的数据源，即便通过获取 E 边界路由器上的 SNMP 数据也会漏报这种异常。



(a)　E:3op　(b)　E:3ob

(c　)E:5op　(d)　E:5ob

图 6　E 5 月 22 日流量数据

6 总结与展望

本文分析现有的异常流量行为检测方法，并根据 NBOS 提供的不同时空的带宽测度选取合适的算法，验证其在大规模高速网络环境中面向不同时空尺度的分析数据源的检测效果。作为研究的起点，我们首先对单位时间的报文数和字节数进行检测。基于指数平滑和基于 Holt-winters 的方法作为时间序列分析方法，能够满足实时性的要求。经过验证，基于 Holt-winters 的检测方法能较好地对网络流量行为预测。最后，将本文的检测方案部署在 CERNET 南京主节点上并对检测结果进行分析。从实验结果看，检测到的异常事件数量偏多，各测度间及相邻时间粒度的检测结果紧密相关。因此后继的工作将从两个角度展开：一是考虑对本文使用的 20 个测度进行关联性分析，并根据多个连续检测点的异常情况进行归并，从而更准确地描述异常事件的等级并报警；二是基于 NBOS 提供的端口分布等流量信息结构数据进行检测，与本文的检测方案相互补充，完善 NBOS 系统异常行为检测的功能。

参考文献

[1]　Leland WE,Taqqu MS,Willinger W,etal.On the self-similar Natrue of Ethernett raffic [J]. ACM SIGCOMM Computer Communication Review, ACM. 1993,23(4):183-193

[2] P.Barford, D.Plonka.Characteristics of Network Traffic Flow Anomalies.In Proceedings of the ACM SIGCOMM Internet Measurement Workshop, Nov, 2001

[3] Lakhina A, Crovella M, Diot C. Diagnosing network-wide traffic anomalies. In: Proc. of the 2004 Conf. on Applications, Technologies, Architectures, and Protocols for Computer Communications. Oregon, 2004. 219−230.

[4] M.ThottanandC.Ji. Proaetive Anomaly Detection Using Distributed Inielligent Agents IEEE Network Volume:12, Issue:5,Sept-Oct.1998,pp21-27.

[5] BillahBaki ,King Maxwell L,Snyder Ralph D,etl. Exponential smoothing model selection for forcecasting.International Jounal of Forcasting, 2006, 22(2):239-247

[6] Brutlag J. Aberrant behavior detect ion in time series for network monitoring [C] .Proceed ings of the USENIX Fourteenth System Administration Conference LISAX IV. California: USENIXA ssoc, 2000: 139- 146

[7] 张维维, 龚俭, 丁伟, 等. NBOS:一个基于流技术的精细化网管系统[A]. CERNET2012 年会[C]. 太原: 太原理工大学出版社, 2012.

[8] Claise B. Specification of the IP Flow Information Export (IPFIX) Protocol for the Exchange of IP Traffic Flow Information. RFC 5101. January 2008.

[9] Brauckhoff D, Tellenbach B, Wagner A, May M, Lakhina A. Impact of packet sampling on anomaly detection metrics. In: Proc. Ofthe 6th ACM SIGCOMM Conf. on Internet Measurement. Rio de Janeriro, 2006. 159−164.

[10] Maxion R A, Frank E. A Case Study of Ethenet Anomalies in a Distributed Computing Environment.IEEE Transaction on Reliability,1990,39(4):433-443.

[11] Denning DE. An intrusion-detection model. IEEE Transactions on Software Engineering, 1987,SE-13:222~232.

[12] Brockwel P J, Davis R A. Introduction to time series and forecasting[M].New York: Springer, 2002: 326- 328

[13] 朱应武，杨家海，张金祥.基于流量信息结构的异常检测[J].软件学报,2010,10(22):2573-2583.

[14] A. Lakhina, M. Crovella, C. Diot, Mining anomalies using traffic feature distributions, in: ACM SIGCOMM, 2005.

[15] HoLL ,Cavuto D J, Papavassiliou S. Adaptive and automated detection of service anomalies in transaction oriented WAN's network analysis,algorithms, implementation,and deployment [J].IEEE Journal of Selected Areas in Communications,2000, 18 (5) : 744 - 757.

Wu Qi, born in 1992, master candidate.Her main research interests include network measurement and network behavior.

DING Wei, born in 1962, Ph.D., professor, Ph.D. supervisor. Her main research interests include computer integrated manufacturing, general search engine, PKI certificate system, remote education under network environment and network behavior.

Background

Due to the fact that network traffic shows certain cyclical characters and stability when it runs normally, once it deviates from the established pattern and breaks the former rules, anomaly is all the more likely. To detect anomaly rapidly and accurately and to respond to anomaly properly is one of the precondition of ensuring the efficient network operation. Whether the network anomaly is detected accurately or not is very important to improve network availability and reliability. Hence, network traffic anomaly detection is becoming a topic of concern.

Scholars have conducted related research for a long time and made fruitful achievements in this field.Many classical anomaly detection algorithms have been well developed. However, few of them have been widely used in engineering, which can be partly ascribed to coarse-grained analysis data source. Analysis of network traffic are mainly based on SNMP data provided by network devices currently. These data are aggregated according to link addresses and cannot reflect information about IP address and port. Under this provision, analysis can only acts on overall traffic of network device port, which will inevitably fail to report some part anomaly events.

NBOS (Network Behavior Observation System) is a network management system intended to provide service quality management and to keep track of security status. Based on flow records, it can provide traffic data in different temporal and spatial scales, offering more desirable data source for anomaly detection. The work of this paper will validate the effects of anomaly detection algorithms on these data.

At first, the network traffic data NBOS can provide is discussed in detail. Further, the paper outlines the network traffic anomalies and introduced related detection methods, which can offer fundamental theory and tactics to real-time network anomaly detection. As a research starting point, we choose the number of bytes and packets per unit time as our detection measures, which is a typical case of volume-based detection. Then we analyze available methods and focus on methods based on the exponential smoothing technology. Specifically, we discuss detection methods based on simple exponential smoothing technology and Holt-winters seasonal model. As they are both prediction-based, we compare their ability to describe normal network traffic behavior in virtue of NBOS by computing the deviation between actual traffic and prediction. The experiment results show that Holt-winters model has proved to do better in modeling normal network traffic. Accordingly, the method is deployed at one node of CERNET. Looking into the detection results, we can know that based on traffic data provided by NBOS, network traffic anomaly detection can work well and help network administrators manage the network more effectively.