文章编号：1008-0562(2013)04-0544-05

基于时间特征的网络流量异常检测

刘仁山，孟祥宏

（呼伦贝尔学院 计算机科学与技术学院，内蒙古 呼伦贝尔 021008）

摘　要：为了解决传统网络管理方法不能适应网络复杂性、不能准确刻画网络异常行为的问题.采用一种基于时间特征的网络流量异常检测模型，研究分析网络流量的变化规律.利用指数平滑预测算法对未来网络流量进行预测，利用中心极限定理并结合实际经验确定动态的网络流量阈值，对当前和未来的网络流量异常进行检测.研究结果表明：当网络流量发生异常时，该模型能够进行有效的检测，能准确地描述网络的运行状况.该算法提高了网络流量检测的智能性，具有较高的实用价值.

关键词：网络管理；网络流量；时间特征；异常检测；流量阈值；指数平滑算法；中心极限定理；流量预测

中图分类号：TP 393　　**文献标志码：**A

Anomaly detection of network traffic based on time characteristics

LIU Renshan, MENG Xianghong

(College of Computer and Science, Hulunber College, Hulunber 021008, China)

Abstract: In order to overcome the problems associated with the traditional network management method, which is unable to meet the requirements of current network complexity and accurately describe the network abnormal behavior, the paper investigates and analyzes the network traffic variations using a network abnormal traffic detection model based on time characteristics. Exponential smoothing algorithm is used to forecast the future network traffic, a dynamic network traffic threshold is determined with central limit theorem and practical experience, therefore, the current and future network abnormal traffic can be detected. The study results show that the traffic model can effectively detect the abnormal traffic and accurately describe the network operation status. This algorithm improves the intelligence of network traffic detection and has a very high practical value.

Key words: network management; network traffic; time characteristics; anomaly detection; traffic threshold; exponential smoothing algorithm; central limit theorem; traffic forecast

0　引　言

随着通信网络技术的发展，网络复杂性不断增加，传统的网络管理方法已经不能适应发展的需要，网络管理员需要借助智能化的网络管理方法对网络进行有效的管理.流量预测和流量异常检测是网络管理的有效手段[1-2]，文献[3]比较了基于时间序列的流量统计模型、泊松模型和自相似模型，提出了基于时间序列流量统计模型具有准确性较高的观点.基于此针对校园网流量的特点提出基于时间特征的网络流量异常检测方法，结合自适应阈值对流量异常进行检测.

预测就是根据历史数据的波动特性和变化趋势，采用合适的预测方法对未来网络流量变化趋势做出预测，预测方法可分为定性方法和定量方法，定性方法是基于专家判断的预测方法；而时间序列预测方法和因果预测法则属于定量方法，时间序列预测法分为移动平均法、指数平滑法、趋势预测法、HOLT 预测法等，多数预测方法都把时间序列分为 4 种成分[4]，即趋势成分、季节成分、循环成分和不规则成分.对于没有明显趋势、季节和循环成分的时间序列可以用移动平均法和指数平滑法进行预测；对于有长期线性趋势的时间序列，可以用趋势预测法进行预测；对既有线性趋势又有季节成分的时间序列可以用 HOLT 预测法进行预测.因果预测法主要是回归分析，采用统计的方法建立变量之间的数学方程式，从而建立预测模型，

根据自变量个数的多少回归模型可以分为一元回归模型和多元回归模型；根据回归模型是否线

收稿日期：2012-08-30
基金项目：内蒙古自然科学基金资助项目（2011BS0905）；国家社会科学基金资助项目（11XTQ009）；呼伦贝尔学院科学研究重点项目（YJZDZC201202）
作者简介：刘仁山（1974-），男，内蒙古 呼伦贝尔人，硕士，副教授，主要从事网络管理、网络安全等方面的研究.　本文编校：朱艳华

性，回归模型可以分为线性模型、指数模型、修正指数曲线模型、逻辑斯谛曲线模型（皮尔曲线模型）和非线性模型[5]等.

通过对校园网流量散点图的分析，发现在一定时间序列中网络流量围绕一个水平上下波动，基于这种特点并结合指数平滑预测算法的特性，本文采用指数平滑法进行网络流量预测，同时为了能够及时发现网络的异常流量，利用置信区间法设置具有自适应的流量阈值，该方法建立的数学模型不仅可以对未来的网络流量做出预测，而且能够对已经发生或即将发生的网络流量异常进行报警以备分析之用.

1　网络流量模型算法

1.1　指数平滑算法

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，是一种性能优良、适应性强的方法，在各个方面都有着广泛的应用[6-7].其作用主要体现在两个方面：一是用于预测，二是用于修匀历史数据，以测定时间数列的长期趋势.指数平滑法具有显著的特点，它有效地利用了全部历史数据且操作简单易行.

指数平滑法是一种加权的移动平均法，它通过将过去所有时期的观测值的加权移动平均数作为下一时期的预测值

$$F_{t+1}=Y_t+\alpha(1-\alpha)Y_{t-1}+\alpha(1-\alpha)^2Y_{t-2}+\cdots,$$

式中，α称为平滑系数，$0\leq\alpha\leq1$.

证明上式中的权数随着期数的前推而减少，并且权数之和等于 1，

$$\alpha\succ\alpha(1-\alpha)\succ\alpha(1-\alpha)^2\succ\alpha(1-\alpha)^3\succ\cdots,$$
$$\alpha+\alpha(1-\alpha)+\alpha(1-\alpha)^2+\alpha(1-\alpha)^3+\cdots=1.$$

因此指数平滑法在预测时更重视近期的数据，α 值越大，近期观测值对预测的影响越大，α 值越小，近期观测值对预测的影响越小.

预测值可变换成

$$F_{t+1}=\alpha Y_t+(1-\alpha)F_t,\qquad(1)$$

式中，F_{t+1}是 t+1 时期的预测值；Y_t是 t 时期的观测值.

α 的选取值对于预测的准确性至关重要，如果时间序列有较大的随机波动，说明多数预测误差是由随机因素引起的，此时应选择较小的平滑系数，这样可以减少由随机因素引起的预测误差对下期预测值的影响.反之，如果时间序列有较小的随机波动，则应选择较大的平滑系数，这样做的好处是可以迅速调整预测值.在本文中，采用均方误差（MSE）极小的原则来确定平滑系数.每一个 α 值确定一个指数平滑预测序列，对每个预测序列求均方误差，取均方误差最小的 α 值进行平滑指数预测，公式为

$$MSE=\frac{1}{n}\sum_{t=1}^{n}(Y_t-F_t)^2\quad.\qquad(2)$$

1.2　置信区间算法

为了了解网络流量是否异常，根据网络流量的历史值采用置信区间法计算网络流量的阈值，根据中心极限定理，如果所研究的随机变量 X 可以表示成很多个独立的随机变量 X_1，X_2，…，X_n之和，只要每个 X_i（i = 1，2，…，n）对 X 只起微小的作用，不管这些 X_i服从什么分布，在 n 比较大的情况下，就可以认为 X 服从正态分布.由于网络流量观测值都是独立的随机变量，因此这些观测值可以使用该定理进行估计.当 $n\prec30$ 时，可以证明

$$\frac{\overline{X}-\mu}{\frac{S}{\sqrt{n}}}\sim t(n-1)\quad.\qquad(3)$$

式（3）表示随机变量 $\frac{\overline{X}-\mu}{\frac{S}{\sqrt{n}}}$服从自由度为 n-1 的 t 分布，将式（3）转换为 μ 的置信区间的形式，即

$$P\left(-t_{\frac{\alpha}{2}}(n-1)\prec\frac{\overline{X}-\mu}{\frac{S}{\sqrt{n}}}\prec t_{\frac{\alpha}{2}}(n-1)\right)=1-\alpha,$$

得

$$P\left(\overline{X}-t_{\frac{\alpha}{2}}(n-1)\frac{S}{\sqrt{n}}\prec\mu\prec\overline{X}+t_{\frac{\alpha}{2}}(n-1)\frac{S}{\sqrt{n}}\right)=1-\alpha.\qquad(4)$$

简写为

$$\left(\overline{X}\pm t_{\frac{\alpha}{2}}(n-1)\frac{S}{\sqrt{n}}\right),\qquad(5)$$

式中，$t_{\frac{\alpha}{2}}$可以根据正态分布表查得，样本标准差

$S=\sqrt{\dfrac{\sum\left(X-\overline{X}\right)^2}{n-1}}$，样本均值 $\overline{X}=\dfrac{\sum X}{n}$.

2 网络流量模型过程分析

网络流量模型首先计算预测值，将每天同一时刻的网络流量值作为预测的历史样本，估算下一天同一时刻的网络流量，周一至周五的流量具有相似性，可作为估算工作日流量的样本；而周六和周日的流量可作为估算休息日的历史样本.若样本的时间跨距为 1 d，可利用前 20 个工作日的样本估算第 21 个工作日的流量，若样本的时间跨距为 2 天，可利用前 40 个工作日中的 20 个样本估算第 22 个工作日的流量，因此可以对未来一个时间段网络流量进行预测.

完善的网络性能管理，不仅仅要进行网络流量预测，还要判断网络流量是否在正常的范围之内，需要设置合理的流量阈值，如果阈值设置太小，可能导致虚假的网络流量异常频繁发生；如果阈值设置过大，可能导致网络流量异常被遗漏.基于这种考虑，本文设置的自适应阈值是根据网络实际流量总体趋势的不同，在每天的不同时刻采用不同的阈值.实现方法是采用前 20 个样本的历史数据，根据中心极限定理并结合文献[8-10]确定网络流量阈值，最近观测值的变化会动态刷新流量阈值，这样可以更符合检测的实际需要，克服固定阈值无法正确描述网络流量动态性的缺点.

网络流量模型的实现过程如下：

步骤 1 从 SNMP 管理信息库 MIB 中定时读取接口组变量：ifInOctets(B/s), ifOutOctets(B/s)；IP 组变量：ipForwDatagrams(packet/s), pInReceives (packet/s).它们分别表示设备接口每秒接收到和发送的字节数、设备每秒接收和转发的数据包数，采集到上述数据可以计算得到某一个时间段的数据传输速率和数据包传输速率.

步骤 2 采用滑动窗口模型，新的 N（滑动窗口大小）个数据指的是最近 N 天同一时刻的流量值，假设 N 取 20，有 $S(n)=T_{(i,j)}$，式中，$S(n)$是滑动窗口中的某个流量值，$n=1,2,\cdots,N$，$T_{(i,j)}$表示第 i 天第 j 时刻网络流量值.按时间排列 $S(1),S(2),\ldots,S(N)$, 随着时间的推移，滑动窗口向前移动.

步骤 3 利用指数平滑法预测流量时，选取合适平滑系数 α 至关重要，设 $\alpha_{(k)}=0.1k$, $k=1,2,\cdots,9$. 对滑动窗口中 $S(n)$序列利用平滑系数 $\alpha(k)$求出预测序列，共九个预测序列，求均方误差 MSE 最小的序列对应的 α 值，这个 α 值就是最优的指数平滑系数，随着滑动窗口的移动 α 值也是动态变化的.

步骤 4 选取最优指数平滑系数 α，利用式（1）对滑动窗口的下一个值进行预测，得到预测值 $Y_{(m,j)}$，表示第 m 天第 j 时刻网络流量预测值，其中 $m=i+N$.

步骤 5 按照中心极限定理确定流量置信区间，对滑动窗口中 $S(n)$序列利用式（5）计算流量置信区间为

$$\left(\overline{X}-t_{\frac{\alpha}{2}}(n-1)\frac{S}{\sqrt{n}},\overline{X}+t_{\frac{\alpha}{2}}(n-1)\frac{S}{\sqrt{n}}\right),$$

式中，$\overline{X}$ 是滑动窗口中 $S(n)$的平均值；$n\leq N$；S 是滑动窗口中 $S(n)$的均方差.

步骤 6 为了防止虚警和漏警的发生，根据经验将网络流量阈值定义为

$\left(\overline{X}-3\times t_{\frac{\alpha}{2}}(n-1)\frac{S}{\sqrt{n}},\overline{X}+3\times t_{\frac{\alpha}{2}}(n-1)\frac{S}{\sqrt{n}}\right)$，得到上边界 $R_{(m,j)}=\overline{X}+3\times t_{\frac{\alpha}{2}}(n-1)\frac{S}{\sqrt{n}}$，下边界 $L_{(m,j)}=\overline{X}+3\times t_{\frac{\alpha}{2}}(n-1)\frac{S}{\sqrt{n}}$，式中，$R_{(m,j)}$表示第 m 天 j 时刻的上边界；$L_{(m,j)}$表示第 m 天 j 时刻的下边界，根据所乘系数的不同可以设置不同的流量阈值.

在观测网络流量时以一个滑动窗口为参照，检测该滑动窗口下一个相邻的网络流量值是否发生异常变化，根据检测到的点是否落到自适应阈值上下的置信区间内判断当前点是否异常，判断规则如下：

当 $R_{(m,j)}>Y_{(m,j)}>L_{(m,j)}$时，预测值判断为正常；

当 $R_{(m,j)}>T_{(m,j)}>L_{(m,j)}$时，实际观测值判断为正常；

当 $Y_{(m,j)}>R_{(m,j)}$ 或 $Y_{(m,j)}<L_{(m,j)}$时，预测值判断为异常；

当 $Y_{(m,j)}>T_{(m,j)}$ 或 $T_{(m,j)}<L_{(m,j)}$时，实际观测值判断为异常.

3 数据仿真

为验证数学模型的可行性和准确性，以时间间隔为 1 h 采集交换机接口 MIB 数据，采集时间为

2012 年 4 月 2 日至 4 月 27 日, 采集每天 7:00-12:00 的数据，这样每天同一时刻的数据形成一个滑动窗口样本（这里除去休息日，因为休息日的网络流量与工作日流量不具有相似性，对于休息日流量的预测可以采集休息日的流量样本进行计算）.采集的交换机是校园网的核心交换机 S9505，流量采集和预测任务一台 PC 机 NMS 来完成.采集数据见表 1.

表 1　滑动窗口中的流量样本值（单位：Mb）

Tab.1　traffic sample value of sliding window (unit: Mb)

时刻	日　期																			
	2012-04-02	2012-04-03	2012-04-04	2012-04-05	2012-04-06	2012-04-09	2012-04-10	2012-04-11	2012-04-12	2012-04-13	2012-04-16	2012-04-17	2012-04-18	2012-04-19	2012-04-20	2012-04-23	2012-04-24	2012-04-25	2012-04-26	2012-04-27
	S(1)	*S*(2)	*S*(3)	*S*(4)	*S*(5)	*S*(6)	*S*(7)	*S*(8)	*S*(9)	*S*(10)	*S*(11)	*S*(12)	*S*(13)	*S*(14)	*S*(15)	*S*(16)	*S*(17)	*S*(18)	*S*(19)	*S*(20)
07:00	214	223	235	212	245	234	213	223	234	231	239	227	238	217	249	231	210	223	233	245
08:00	235	246	256	267	277	267	267	269	258	254	278	276	247	241	242	243	256	237	268	257
09:00	323	345	332	315	328	341	338	346	341	348	326	349	341	302	309	343	341	346	322	312
10:00	378	408	365	389	392	378	383	375	398	389	361	398	408	402	389	372	381	365	384	375
11:00	398	456	423	412	399	400	447	456	438	446	423	433	412	436	456	423	432	439	454	427
12:00	432	398	427	436	416	431	438	425	421	397	392	403	409	417	414	426	437	408	409	398

利用步骤 3 计算每个滑动窗口对应的平滑系数α，利用步骤 4 预测下一个相邻值（4 月 30 日的网络流量）；利用步骤 5 和步骤 6 计算网络流量阈值，表 2 为 4 月 30 日流量预测值、流量实际值和流量阈值，通过分析 4 月 30 日 11:00 的流量实际值出现了异常, 其它时间段网络流量预测和实际值均处于正常范围之内.滑动窗口向前移动一天, 5 月 1 日的流量情况见表 3，11:00 预测值和实际值都超出了正常的阈值范围, 流量出现异常, 网络管理员根据预警信息分析网络性能，查找原因.

表 2　4 月 30 日流量预测值、实际值和阈值

Tab.2　traffic prediction value、real value and threshold on April 30

时刻	平滑系数	流量预测值/Mb	流量实际值/Mb	阈值上边界	阈值上边界
07:00	0.5	236	243	252	205
08:00	0.7	258	245	282	232
09:00	0.6	319	344	361	304
10:00	0.5	377	382	415	354
11:00	0.8	432	278	463	398
12:00	0.8	400	408	448	385

表 3　5 月 1 日流量预测值、实际值和阈值

Tab.3　traffic prediction value、real value and threshold on May 1

时刻	平滑系数	流量预测值/Mb	流量实际值/Mb	阈值上边界	阈值上边界
07:00	0.5	240	251	254	207
08:00	0.7	249	273	283	233
09:00	0.6	334	321	362	305
10:00	0.5	379	363	415	354
11:00	0.8	469	484	467	402
12:00	0.8	406	432	447	384

进行网络流量预测的关键是历史样本数据的收集，收集的数据越多计算就越准确，但是计算量就越大，本文定义的滑动窗口收集 20 个样本，在滑动窗口没有滑动之前，只能采用有限的样本进行预测和计算阈值，预测准确性不高，随着滑动窗口的移动，预测值和阈值会越来越准确，见图 1，在 4 月 3 日至 4 月 10 日之间，网络阈值出现的偏差较大，随着采集样本的增加，阈值趋于正常，在 4 月 30 日和 5 月 1 日网络流量高于阈值上边界，网络流量出现异常.

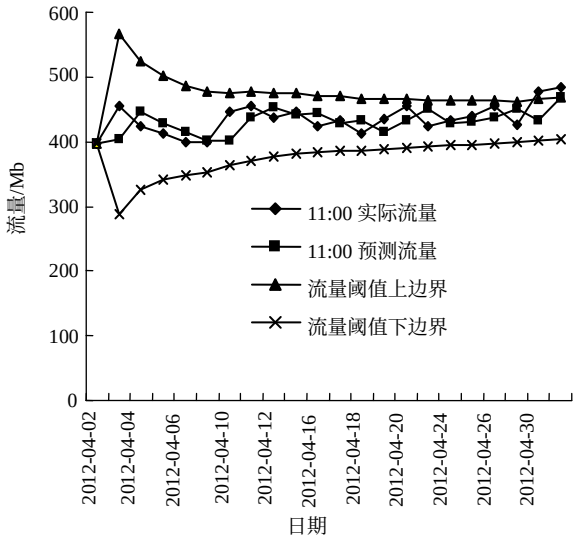

图 1　网络流量分析

Fig.1　network traffic analysis diagram

4　结　论

提出了基于时间特征的网络流量异常检测方

法，该方法可根据网络流量实时调整模型参数和检测统计量，具有如下优点：

（1）算法的计算过程简单，计算量少，灵活易于实现，滑动窗口大小可以变化，预测的时间单元可以是 1 h、30 min 等，可以预测未来几天的网络流量.

（2）预测流量和自适应阈值对于网络流量监控、性能分析和峰值监测有很好的参考使用价值.

（3）这种方法可以反映出流量的时间趋势及其变化，而不只是使用统计的方法来反映流量的趋势变化.

（4）实验表明该方法能够实时检测，网络流量预测准确率较高，能帮助网络管理员对网络性能进行分析以及对故障进行诊断，减少网络的拥挤和不可访问性，提高了网络管理的智能性.

参考文献：

[1] 刘渊,马汝辉,林星.基于QPSO小波神经网络的网络异常检测[J].辽宁工程技术大学学报:自然科学版,2011,28(2):261-264.
Liu Yuan,Ma Ruhui,Lin Xing.Network anomaly detection for wavelet neural network based on QPSO [J].Journal of Liaoning Technical University:Natural Science,2011,28(2):261-264.

[2] 张国权,李文立.基于混合互信息的决策树入侵检测[J].辽宁工程技术大学学报:自然科学版,2011,28(2):273-276.
Zhang Guoquan,Li Wenli.Intrusion detection based on decision tree with mutual information[J].Journal of Liaoning Technical University:Natural Science,2011,28(2):273-276.

[3] 叶新铭,王斌.基于时间特征的网络流量预测模型[J].计算机科学,2005,32(7):34-37.
Ye Xinming,Wang Bin.A network traffic forecast model based on time character[J].Computer Science, 2005,32(7):34-37.

[4] 汪淼,郑舒婷.基于 ARIMA 模型的中国消费者价格指数时间序列分析[J].辽宁工程技术大学学报:自然科学版,2010,29(S1):130-132.
Wang Miao,Zheng Suting.Time sequence analysis of CIP based on ARIMA model[J].Journal of Liaoning Technical University:Natural Science,2010,29(S1):130-132.

[5] 董梦丽,杨庚.网络流量预测方法[J].计算机工程,2011,37(16):98-100.
Dong Mengli, Yang Geng. Methods of network traffic prediction[J]. Computer Engineering,2011,37(16):98-100.

[6] Wood D J, Ormsbee L E. Supply identification for distribution systems[J]. Joural AWWA, 1989(7): 74-80.

[7] Joseph J, LaViola Jr. An experiment comparing double exponential smoothing and kalman filter-based predictive track in galgorithms [EB/OL]. http://www.cs.brown.edu/people/jjl/pubs/vr2003 laviola.pdf, 2005(5):5.

[8] 梁昇,肖宗水.基于统计的网络流量异常检测模型[J].计算机工程,2007,31(24):123-125.
Liang Sheng, Xiao Zongshui. Anomaly detection model of network traffic based on statistics[J]. Computer Engineering, 2007,31(24):123-125.

[9] 曹敏,程东年.基于自适应阈值的网络流量异常检测算法[J].计算机工程,2009,35(19):164-166.
Cao Min,Cheng Dongnian. Network traffic abnormality detection algorithm based on self-adaptive threshold[J]. Computer Engineering, 2009,35(19):164-166.

[10] 邹柏贤. 一种网络异常实时检测方法[J].计算机学报,2003,26(8): 940-945.
Zou Boxian. A real-time detection method for network traffic anomalie[J]. Chinese Journal of Computers, 2003, 26(8) :940-945.