一种新的网络异常流量检测模型*)

涂旭平　金　海　何丽莉　杨志玲　陶智飞
(华中科技大学计算机科学与技术学院　武汉 430074)

摘　要　网络的异常流量检测是通过比较系统或用户的实际行为模式与正常行为模式之间的区别来检测入侵,目前的异常流量检测系统没有均衡考虑检测实时性与检测可信度之间的矛盾,本文提出了一种可调 Chi-Square T^2 的网络异常流量检测模型,根据测度的可信度不同,设置不同的可信系数,通过调整可信系数,使总的异常值可以反映测度的可信度,从而提高了检测效率。

关键词　流量检测,异常检测,测度系数

A New Model of Anomaly Network Flow Detection

TU Xu-Ping　JIN Hai　HE Li-Li　YANG Zhi-Ling　TAO Zhi-Fei
(Huazhong University of Science and Technology, Wuhan430074)

Abstract　The main methods to detect anomaly network flow are through comparison between the system's actual action pattern and normal action pattern. There is a trade-off between real-time and veracity in anomaly network flow detection, this paper presents a new model -adjustable Chi-Square T^2 model, according to a set of trusty coefficient. Through adjusting these coefficients, a trusty measure can contribute more value to the total anomaly value with a high factor, this will observably improve detection rate.

Keywords　Flow detect, Anomaly detect, Measure coefficient

1　前言

异常流量检测是目前 IDS(入侵检测系统)研究的一个重要分支,这种检测方法通过建立系统或用户的正常行为模式库,比较系统或用户的实际行为模式和正常行为模式之间的区别来检测入侵,其特点是不需要过多地了解被保护系统的缺陷,具有较强的适应性,能够检测出未知入侵,但存在虚警概率高的缺点。其核心问题是如何实现流量正常行为的描述、检测的实时性、获得信息的全面性和反应的灵敏性,因而使系统设计和实现难度加大。因此面向网络的实时安全监测系统是当前研究的一个热点。实时异常检测的前提是能够实时,对大规模高速网络流量进行异常检测首先要面临高速流量载荷问题,由于测度、分析和存储等计算机资源的限制,无法实现全网络现流量的实时检测,因此,抽样测度技术成为高速网络流量测度的研究重点。网络异常流量检测主要针对产生异常流量的攻击行为,如端口扫描、flood 型 DDoS 攻击和蠕虫等。Medhi Nassehi[1]等利用马尔可夫模型检测异常,Luca Deri[2]等实现了一个类似网络管理工具的网络异常检测系统,Polly Huang[3]等用小波的方法检测网络的性能问题;利用小波的方法检测网络异常的还有 Aquino 和 Barria[4]。中国科学院计算技术研究所的邹柏贤和李忠诚[5]实现了 AR 自回归模型检测异常,通过维护一个 N 大小的窗口,求得一个统计量的估计值,当获得了第 $N+1$ 个数据的时候,窗口向前滑动,重新计算统计量 R 的值。通过比较 R 值来发现异常。本文在 T^2 的基础上提出了一种新的异常检测模型;通过调整 Chi-SquareT^2 系数来提高检测率。

2　可调 Chi-Square T^2 检测模型

可调 Chi-Square T^2 检测模型是在 T^2 模型及加权平方和模型的基础上发展而来,本节将先对后面两个基础模型进行介绍,然后提出可调 Chi-Square T^2 检测模型。

2.1　模型理论基础

1) T^2 模型:统计值 T^2 是对多个测度的综合评价,设有 n 个可以表征异常的测度,记为 X_i,$1\leqslant i\leqslant n$,测度 X_i 与 X_j 之间的相关性记为 C_{ij},$1\leqslant i,j\leqslant n$,统计值 T^2 定义为:

$$T^2=(X_1,X_2,\cdots,X_n)C^{-1}(X_1,X_2,\cdots,X_n)' \quad (1)$$

C^{-1} 表示向量$(X_1,X_2,\cdots,X_n)$相关矩阵的逆矩阵,$(X_1,X_2,\cdots,X_n)'$ 表示该向量的转置向量。当测度值 X_i 互不相关时,T^2 简化为 $T^2=X_1^2+X_2^2+\cdots+X_n^2$,由于很难确定 C_{ij},因此公式 T^2 很难实现。当相关系数不为零时,T^2 则为一个考虑向量 X_i 相关性的复杂函数。

2)加权平方和模型:在式(1)中,由于难于确定 C^{-1},采用近似公式进行处理。将 C_{ij} 的非对角线上的元素都设定为 0,统计值 T^2 简化为:

$$T^2=a_1X_1^2+a_2X_2^2+\cdots+a_nX_n^2 \quad (2)$$

在式(2)中,a_i($1\leqslant i\leqslant n$)为管理员指定的正系数。根据 T^2 的定义可知每个测度值 X_i 对于 T^2 贡献的大小。而安全管理员可通过提高系数的值来提高 X_i 的权值。当 T^2 取加

*)武汉市重点攻关资助项目(20031003027),湖北省自然科学基金资助项目(2001ABA001)。涂旭平　博士生,研究方向为并行与分布式软件、网络安全与网格安全。金　海　博士生导师,研究方向为计算机系统结构、并行与分布处理、集群与网络计算、高性能外存储系统、无形计算与无形存储、网络安全。

权平方和时，它将不再对各个 X_i 之间的相关性敏感了。IDES[6]系统将重新定义为：

$$T^2=\sum_{1}^{n}a_iX_i^2+\sum_{1}^{n}a_{ij}h(X_i,X_j,C_{ij}) \quad (3)$$

在式(3)中，$a_i(1\leqslant i\leqslant n)$为管理员指定的正系数，$h(X_i,X_j,C_{ij})$是关于 X_i，X_j 及其相关矩阵 C_{ij} 的函数，该函数当 X_i，X_j 与其历史相关性异常时，取较大值，因而新的 T^2 定义既可以解决前面所述的困难，又反映出测度值之间的协方差的变化，这个模型同样面临系数 $h(X_i,X_j,C_{ij})$难以确定的问题。

2.2 可调 Chi-Square T^2 模型的建立

Chi-Square 模型中统计量 T^2 定义为 $T^2=\sum_{1}^{n}\frac{(X_i-\overline{X}_i)^2}{\overline{X}_i}$[7]，(原公式为 $X^2=\sum_{1}^{n}\frac{(X_i-\overline{X}_i)}{\overline{X}_i}$，为了叙述方便，$X^2$ 在不改变意义的情况下替换为 T^2)其中 $\overline{X}_i$ 是 X_i 的均值，根据中心极限定理，当 T^2 的样本足够大时，T^2 将服从正态分布，故区间$[\mu-Z_{\alpha/2}\delta,\mu+Z_{\alpha/2}\delta]$覆盖 T^2 总体的百分比约为$(1-\alpha)\times100\%$，其中 μ,σ 分别为 T^2 的总体均值和总体方差，$Z_{\alpha/2}$为阈值因子，一般设置为 3[8,9]，通过对 T^2 的多次观察得到的样本均值 $\overline{T}^2$ 和样本方差 $S_{\overline{X}}^2$ 估计，以样本均值和样本方差分别代替总体均值和总体方差，区间$[\mu-Z_{\alpha/2}\delta,\mu+Z_{\alpha/2}\delta]$变为$[\overline{X}^2-3S_{\overline{X}}^2,\overline{X}^2+3S_{\overline{X}}^2]$，由于本文是检测显著大的 T^2，因此只需考虑该区间的上限，即当检测到 $T^2>\overline{X}^2+3S_{\overline{X}}^2$ 时，可认为系统发生了异常。

由于不同的测度计算出的$\frac{(K_i-\overline{K}_i)^2}{\overline{K}_i}$与$\frac{(K_j-\overline{K}_j)^2}{\overline{K}_j}$可能处于不同的数量级别，如$\frac{(K_i-\overline{K}_i)^2}{\overline{K}_i}$计算出的值分布在某常数 ValueA 附近，而$\frac{(K_j-\overline{K}_j)^2}{\overline{K}_j}$计算出的值分布在另一常数 ValueB 附近，如果 ValueA 远大于 ValueB，则 $T^2=\sum_{1}^{2}\frac{(K_i-\overline{K}_i)^2}{\overline{K}_i}$的值将基本上由$\frac{(K_i-\overline{K}_i)^2}{\overline{K}_i}$控制，而受$\frac{(K_j-\overline{K}_j)^2}{\overline{K}_j}$的影响很小。可引入一个系数 d_i 来平衡这种差异。综合 2.2 节的模型，该系数表示了不同的测度对于 T^2 的影响，得到一种更加综合有效的可调 Chi-SquareT^2 模型：$T^2=\sum_{1}^{2}d_i\frac{(K_i-\overline{K}_i)^2}{\overline{K}_i}$，其中 d_i 为不同的测度确定的系数。

2.3 模型系数的确定原则

在上节提出的可调 Chi-Square T^2 模型基础上，根据检测可信度需要，确定如下的模型系数：

1) 与测度所关联的对象的普遍性相关　测度所关联的对象越普遍，则系数应该越大。例如 TCP 的流量在所有的网络流量中是占很大的比例的，如果这个测度是 TCP 相关的，那么它的值比 UDP、ICMP 或其他非常用协议相关的测度的值要大。

2) 与$\frac{(K_i-\overline{K}_i)^2}{\overline{K}_i}$值的规模成反比　对于大多数的测度，假设$\frac{(K_i-\overline{K}_i)^2}{\overline{K}_i}$的值分布在 10 附近，另假定存在 10 个测度，编号为 1～10，对于其他测度的值分布在 100 附近，假定存在两个，编号为 11，12，选定一个基准 10，调整这两个测度与前 10 个测度在同一等级 10。如果不受原则 1)的影响，则 $d_1=d_2=\cdots=d_{10}=10r$，$d_{11}=d_{12}=r$，其中 $r>0$ 为一个小的等分值，且满足 $\sum_{1}^{12}d_i=1$。在实际确定系数时，需综合考虑这两条原则。

2.4 模型测度的选定原则

测度的选取直接决定模型的效果，本文采用了基于包个数的模型测度及基于时间序列的模型测度两种方法。

2.4.1 基于包个数的模型测度　每次在获取固定个数的包后，分析检测是否存在异常，通过与以前的训练阈值相比较，判断有无异常发生。本方法共采用两个测度，包的大小、包的时间间隔。基于这两种测度，分别检测 TCP 包数据，UDP 包数据，ICMP 包数据，IP 数据判断是否存在异常。可获知在$[\overline{X}^2-3S_{\overline{X}}^2,\overline{X}^2+3S_{\overline{X}}^2]$的阈值区间内，误警率非常高，本文通过对区间进行修正，当区间调整为$[0,\overline{X}^2+PS_{\overline{X}}^2]$，且 P∈[20,24]时误警率有显著的降低，其中 P 为阈值因子。测度选择方法中存在一个问题：计算的条件是获得了固定个数的包(设为 M)，如果在获得了 $M-1$ 个包后，长时间无法获得第 M 个包时，系统就无法进行计算。但这种可能对系统影响不大，因为通常选择的 M 不是很大，而且只有在系统轻负荷的时候才会出现这种情况，但另一方面，如果系统负荷较轻，可假设发生攻击的可能性更低，因为概率方式的检测，对密集性的异常行为更敏感。

2.4.2 基于时间序列的模型测度　在一个固定小的时间段内，分别统计不同类型包的平均个数，平均大小，以及时间间隔。本测度选择方法分析 TCP，UDP，ICMP 包在一分钟中平均个数，平均大小，以及所有 IP 包平均时间间隔。该方法共包括 7 个测度，分别为：TCPPacketSize(TCP 包平均大小)，UDPPacketSize(UDP 包平均大小)，ICMPPacketSize(ICMP 包平均大小)，TCPPacketNumber(TCP 包平均个数)，UDPPacketNumber(UDP 包平均个数)，ICMPPacketNumber(ICMP 包平均个数)，IPPacketIntervel(IP 包平均时间间隔)。由于 UDP，ICMP，TCP 包在一天的后期发送间隔时间变得非常大，使得时间间隔的平均值与方差明显偏大，不适合使用正态分布的模型，因此将三种包的时间间隔测度综合成一个时间间隔测度来考虑，所以测度缩减为 7 个(实际的包的流量情况见第 3 节模型实现与分析)，把这些测度一起综合考虑求出 T^2。本方法和上述方法相比，其优点是数据的分析次数与被检测系统的流量负载无关，即不管当前被检测系统流量负载大(接收到很多的包)，或者小(接收到少量的包)，需要计算的次数都不变，因为每过一个时间段才计算一次。以下是本测度选取方法的推导过程：

TCP(UDP，ICMP)包的大小，TCP(UDP，ICMP)包的个数，IP 包的平均到达时间，依次为 $X_{0,i}$，$X_{1,i}$，$X_{2,i}$，$X_{3,i}$，$X_{4,i}$，$X_{5,i}$，$X_{6,i}$，其中 $X_{i,j}$ 表示第 $i(i<7)$ 个测度第 $j(j<U-1)$ 次观测值，可得：

$$T_j^2=\sum_{i=1}^{W-1}\frac{(X_{i,j}-\overline{X}_i)^2}{\overline{X}_i},W=7,j=0,1,2\cdots,U-1$$

总体 μ_{T^2} 的(无偏)估计值为：

$$\mu_{T^2}=\overline{T}^2=\frac{1}{U}\sum_{j=0}^{U-1}\sum_{i=0}^{W-1}\frac{(X_{i,j}-\overline{X}_i)^2}{\overline{X}_i}$$

σ_{T^2} 的(无偏)估计值为：

$$\sigma_{T^2}=S_{T^2}=\frac{1}{U-1}\sum_{j=0}^{U-1}(T_j^2-\overline{T}^2)^2=\frac{1}{U-1}(\sum_{j=0}^{U-1}(T_j^2)^2-U$$

$$(\overline{T^2})^2)=\frac{1}{U-1}(\sum_{j=0}^{U-1}(\sum_{i=0}^{W-1}\frac{(X_{i,j}-\overline{X}_i)^2}{\overline{X}_i})^2-U(\overline{T^2})^2),$$

所以 $T_j^2\in[\mu_{T^2}-Z_{a/2}\sigma_{T^2},\mu_{T^2}+Z_{a/2}\sigma_{T^2}]$是正常的，否则就是异常的。本模型也可以将 T^2 中的每项增加一个系数 a_i $(i=0,1,\cdots,W-1)$，得到一个更加合理的模型：

$$\mu_{T^2}=\overline{T^2}=\frac{a_i}{U}\sum_{j=0}^{U-1}\sum_{i=0}^{W-1}\frac{(X_{i,j}-\overline{X}_i)^2}{\overline{X}_i}$$

$$\sigma_{T^2}=S_{T^2}=\frac{a_i}{U-1}(\sum_{j=0}^{U-1}(\sum_{i=0}^{W-1}\frac{(X_{i,j}-\overline{X}_i)^2}{\overline{X}_i})^2-U(\overline{T^2})^2)$$

3 模型实现及分析

本系统使用林肯实验室的数据[10]，只对 inside. tcpdump (inside. tcpdump 是林肯实验室公布的一部分数据名称的前缀，表示其模拟环境中内网部分的 tcpdump 数据，outside. tcpdump 意义类似）的数据进行检测，对 outside. tcpdump 的数据未加分析。对于两种测度的选取方法分别进行了模拟，并且按照第 2.3 节中系数的确定原则，模拟了增加系数后的模型效果。

图 1 显示对于时间间隔，在一天的后期，网络中包的个数明显变少，以至于间隔很大，与一天的前期无法服从同一个分布。将来的模型会将一天分为若干个时间段，对时间段分别计算均值，期望可以获得更好的检测效果。实验数据来自第三周第二天，其它的时间有近视的规律，在此不一一分析。

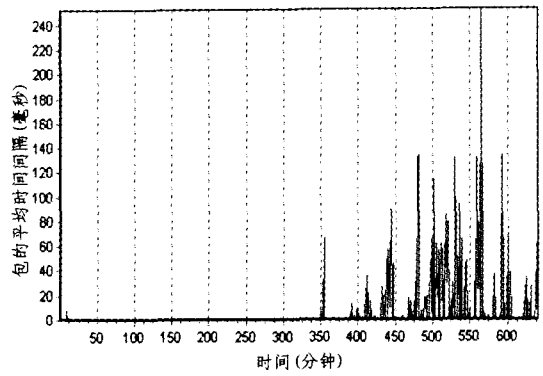

图 1 包到达时间间隔规律

基于包个数的测度模拟 M 表示每 M 个包分析一次，P 表示决定阈值的阈值因子。TCP，UDP，ICMP，IP 分别表示包的类型，如 TCP 表示只对 TCP 包进行分析。表中的五元组数据依次表示为：detections，Alarms，True，False，Times。

表 1 基于包个数的测度模拟

M / P	32						64					
	protocol	Detections	Alarms	True	False	Times	protocol	Detections	Alarms	True	False	Times
20	tcp	2	173	2	171	112	tcpl	1	123	1	122	58
	udp	44	1978	102	1876	676	udp	42	1318	76	1242	352
	icmp	13	323	32	291	305	icmp	8	195	17	178	164
	all	27	1043	62	981	350	all	25	588	42	546	161
21	tcp	2	168	2	166	108	tcp	1	122	1	121	57
	udp	38	1358	70	1288	459	udp	33	927	52	875	253
	icmp	13	323	32	291	305	icmp	8	195	17	178	164
	all	22	735	46	689	246	all	20	415	29	386	117
22	tcp	2	167	2	165	107	tcp	1	120	1	119	56
	udp	31	1063	52	1011	348	udp	24	659	33	626	185
	icmp	13	323	32	291	305	icmp	8	195	17	178	164
	all	20	605	38	567	198	all	14	324	20	304	94
23	tcp	2	165	2	163	105	tcp	1	118	1	117	54
	udp	21	775	35	740	248	udp	15	468	22	446	137
	icmp	13	323	32	291	305	icmp	8	195	17	178	164
	all	13	450	25	425	149	all	10	222	12	210	68

对于 TCP 包的异常检测数据，其中 Times 表示在所有的计算次数中有 Times 次计算发现了异常。由于只能确定是异常，但是无法准确地定位到具体的受攻击者，例如一次异常中，ip 地址为：x. y. z. a 和 x. y. z. b 都出现在目标 ip 地址集合中，因此这两个 ip 地址都被认为受到了攻击，而事实上到 x. y. z. a 的流量可能只是背景流量，因而报警的次数 Alarms 一般远大于 Times。由于存在误警，故 False + True = Alarms，对于多次报警都是同一次攻击的情况，可合并成一次报警，这样得到的报警数可采用 Detections 表示，由表 1 可得，对于相同的 M，当选取 4 种不同的阈值因子（$p=20,21,22,23$）时，ICMP 的检测结果都不变，TCP 协议也有相似的表现。但是对于 UDP 和 ALL 来说，阈值因子 P 增大，显著地影响到检测效果。以（Times－True）/Times 方式计算的误警率大约为 80～90%。从总体效果来说，取 $P=22$ 的效果比较好。对于 UDP 而言检测到的攻击多，误警也高，但是误警率基本不变，稳定在 80～90%。

本实验只是选择性地对 UDP 在阈值因子 $P=22$ 时，进行了系数 d_i 调整，其他几种测度没有进行实验。由于包的到达平均时间间隔与包的平均大小两种测度的重要性大致相当，故期望的系数也应该是(1/2,1/2)，因此预期修改系数意义不大，为了验证检测结果与系数关系不大，另外取两组系数(1/3,2/3)，(2/3,1/3)。由表 2 的数据可知，当测度的重要性相当时，系数取相等为宜。得到的实验数据如表 2 所示。

（下转第 54 页）

但这是在牺牲数据量和加密性能的前提下实现的，如果要传输的数据量很大，并且各个接收者要获取的数据之间有很大的重复度，上述方法将会影响消息传输性能[5]。

在大多数情况下，系统外部的安全性要比系统内部的安全性重要得多，因为系统内部的各个节点都应该是可信的，要牺牲性能来保证不必要的安全性就不合适了，这时就只需要考虑系统外部的安全性，这就是第二种情况。这时就应该改变加密的策略，对整个信息进行加密，发送者用它发布的公钥加密数据，并且保证每个接收者都有发送者的私钥，当消息到达某个接收者时，它就用文章刚开始讲的方法只对属于自己的视图解密，在获得数据后，在将消息发送给下一个接收者。整体加密情况如图 5 所示。

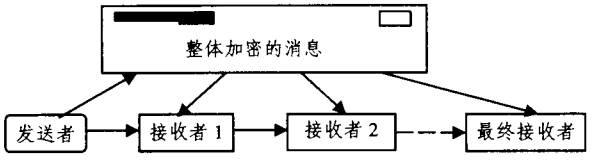

图 5　整体加密示意图

此外在路径的设置上还分为静态和动态两种，当需要在传输的过程中动态改变传输路径时，要考虑防止出现死循环的情况，一旦出现死循环，一部分节点就会永不停止地中转、加密和解密数据，造成传输的失败[6]，不过这不是本文考虑的问题，在此不再讲述。

结论　通过在局域网上进行的实验表明，该加密方法是可行的，在传输的数据量不是特别大的情况下，其性能也是合理的。本文的下一步研究工作将着重研究大数据量情况下的数据冗余和加密性能的问题。

参 考 文 献

1　柴晓路. 为什么需要 Web 服务. IBM developerWorks 专刊，2003
2　King S. CLSSP, threat and Solutions to Web Services Security. Network Security, 2003, 9: 8～11
3　Wahlin D. XML for ASP. NET Developers. SAMS, 2001
4　柴晓路. SOAP Header 扩展：WS-Routing 和 WS-Referral，IBM Corporation, 2001. http://www-900. ibm. com/developerWorks/cn/webservices/ws-soapheadext/part2/index. shtml
5　张勇，冯东雷，陈涵生，白英彩. Internet 密钥交换协议的安全缺陷分析. 软件学报，2002, 13(6)
6　刘怡，李伟琴. 密码协议的分层安全需求及验证. 北京航空航天学学报，2002, 28(5): 589～592

（上接第 39 页）

表 2　基于包个数调整系数后的测试模拟

d_i		32					64				
		Detections	Alarms	True	False	Times	Detections	Alarms	True	False	Times
P=22	1/2,1/2	31	1063	52	1011	348	24	659	33	626	185
	1/3,2/3	30	1163	53	1110	348	24	750	34	716	185
	2/3,1/3	31	1063	52	1011	348	24	803	35	768	186

表 3　调整系数前后(基于包个数)的检测结果比较

阈值因子 $P=22$	系数	detections	Alarms	True	False	False Rate
未使用系数(等价于系数相等，即都为 1/7)	1/7, 1/7, 1/7, 1/7, 1/7, 1/7, 1/7,	55	878	137	741	0.843
调整系数后	8/21, 8/21, 1/21, 1/21, 1/21, 1/21, 1/21	68	802	153	649	0.809

表 3 给出了基于时间序列的模拟检测结果，由表 3 可知，采用时间序列的检测结果要好于基于包的个数，主要原因是基于包的个数的测度选择方法中测度个数太少，不能准确地描述异常。对于基于时间序列的测度方法，测度的个数多，更能准确地反应异常。第二种方法使用了修改过的系数(8/21, 8/21, 1/21, 1/21, 1/21, 1/21, 1/21)，误警率也有所降低，由 0.843 变为 0.809，检测率有所增加，检测到的入侵数由 55 变为 68，可见使用系数调整后系统的检测效果有了显著的改善。

总结　本文提出了一种新的流量入侵检测的模型 $T^2=\sum_1^2 d_i \frac{(K_i-\overline{K}_i)^2}{\overline{K}_i}$，并且着重分析了基于时间序列的实现以及测度的选取方法，并在对数据深入分析的基础上对测度进行了优化，如合并了三种协议的平均时间间隔测度。文中给出了 d_i 的选取方法，有较强的可操作性，但是仍然需要对网络流量有较深入的了解才能真正地确定出一组适合系统的系数 d_i。另外由于测度本身的选取还有待改进，使得总的检测效果还是不够理想。以后的工作将从两个方面展开，第一是对每天的网络流量进行更细致的建模，将不同的时间段采用不同的概率分布模型，第二是对测度的选取进一步研究，优选出更加有效的测度以提高检测率。

参 考 文 献

1　Nassehi M. Anomaly detection for Markov models: [Research report]. IBM Research Division, Zurich Research Laboratory, 8803 Ruschlikon, Switzerland, 1998
2　Deri L, Suin S, Maselli G. Design and Implementation of an anomaly detection System: An empirical approach. http://jake. unipi. it/～deri/ADS. pdf. Aug. 2001
3　Huang P, Feldmann A, Willinger W. A nonintrusive, wavelet-based approach to detecting network performance problems. ACM SIGCOMM internet measurement workshop 2001, San Francisco, USA, Nov. 2001
4　Alarcon-Aquino V, Barria J A. Anomaly detection in communication networks using wavelets. IEE Proc. -Commun. , 2001, 148(6)
5　邹伯贤，李忠诚. 基于 AR 模型的网络异常检测. 微电子学与计算机，2002(12)
6　Denning D E, Neumann P G. Requirements and Model for IDES - A Real-Time Intrusion Detection System. Computer Science Laboratory, SRI International, 1985
7　Ye Nong, Chen Qiang. An Anomaly Detection Technique Based on a Chi-Square Statistic For Detecting Intrusions Into Information Systems. Quality and Reliability Engineering International
8　Montgomery DC. Introduction to Statistical Quality Control. John Wiley & Sons: New York, 2000
9　Banks J. Principles of Quality Control. John Wiley & Sons: New York, 1989
10　MIT Lincoln Laboratory. A public Web site. http://www. ll. mit. edu/IST/ideval/index. html