一种基于统计的网络流量异常检测方法

崔伟兰[1]　尹逊伟[2]　程永强[1]
1　长春工业大学计算机科学与工程学院　吉林　130012
2　中国防卫科技学院　北京　102206

摘要：异常的网络环境往往可以通过观察网络流量的变化而提前发现，因此，如何实时准确地判定流量异常便成为网络异常检测中的重点。本文将自适应滤波理论引入网络流量异常检测，提出一种基于自适应AR模型的网络流量异常检测方法，该方法收敛快、精度高，能降低误报率，从而使异常检测更准确。

关键词：流量模型；异常流量；AR模型；递推最小二乘(RLS)算法

0 引言

本文采用应用时间序列分析技术，通过建立自适应AR模型对网络流量进行检测。AR模型将历史流量收集起来，然后对其进行分析建模，最后确定一个阀值，利用这个阀值对网络流量监测，但普通的AR模型没有自适应更新的能力，模型系数不能随着输入样本自动更新。本文将自适应滤波的理论引入网络流量异常检测中，使用递归最小二乘算法对模型参数进行连续更新、修正，使模型具有动态自适应更新能力，因此利用该方法建立的网络流量模型更精确，从而提高了入侵检测的精确度。

1 自适应AR模型的建立及异常检测

1.1 基本自回归AR模型的建立

AR模型时间序列分析是根据系统观测到的时间序列数据，通过曲线拟合和参数估计来建立数学模型的理论和方法。N阶AR模型的定义为：$x_t=\varphi_1 x_{t-1}+\varphi_2 x_{t-2}+\cdots+\varphi_n x_{t-n}+e_t$。

其中x_t是AR模型的预测信号，N是建立AR模型的阶次，e_t是白噪声，其均值为零，方差为σ_e^2，φ_1，$\varphi_2\cdots\varphi_n$是AR(N)模型的系数。

定义：一时间序列$\{X_t\}$，满足以下条件时称为平稳的时间序列：

(1) $E(X_t)$常数，$\forall t\in T$；

(2) $E(X_t^2)<\infty$，$\forall t\in T$；

(3) $\gamma_x(t+h,t)$和时间t无关。

由于实际的网络流量具有突发性，一般来说网络流量在统计上是不平稳的。因此需要对流量进行适当处理，使其达到或接近稳定。

第1步　用方差分析(ANOVA)的方法对实际网络流量进行预处理，使其达到或接近稳定。

第2步　零均值化。近似平稳的局部时间序列称为滑动时间窗口，窗口大小为$N+1$，$N+1$个数表示为$y_1,y_2,\cdots,y_N,y_{N+1}$用前个数建立AR模型，来判断第$N+1$个数是否异常。为建立AR模型，对前个数进行零均值化，设$\bar{y}$表示$y_1,y_2,\cdots,y_N$的平均值。

令$x_t=y_t-\bar{y}$，$t=1,2,\cdots,N,N+1$

则$x_1,x_2,\cdots,x_N,x_{N+1}$就是零均值时间序列。

第3步　拟合AR模型。我们取常用二阶自回归模型AR(2)，窗口大小N为20，AR(2)的模型是

$$x_t=\varphi_1 x_{t-1}+\varphi_2 x_{t-2}+e_t$$

用时间序列$x_1,x_2,\cdots,x_N$来估计AR(2)模型参数φ_1，φ_2及σ_e^2，记

$$Y=\begin{bmatrix}x_3\\x_4\\\vdots\\x_N\end{bmatrix}\qquad X=\begin{bmatrix}x_2 & x_1\\x_3 & x_2\\\vdots & \vdots\\x_{N-1} & x_{N-2}\end{bmatrix}\tag{1}$$

并且有　　$E(X_t^2)<\infty$，$\forall t\in T$；　　(2)

$$x_t=y_t-\bar{y}\tag{3}$$

由上述公式可估计出AR的参数及白噪声。

这是普通的AR模型，没有考虑模型的自适应更新能力，本文将自适应滤波的理论引入AR模型建模，使AR参数能进行自适应更新。

1.2 参数的自适应更新

本文引入自适应滤波的理论中的递推最小二乘(RLS)算法，其基本原理是：一个n阶线性预报模型的参数向量$\hat{\varphi}_n(t)$是观察区间$1\leqslant i\leqslant t$内在最小二乘意义下最优解。在i时刻，预报误差为

$$\varepsilon(i)=x(i)-\hat{\varphi}_n^T X_n(i-1)\quad 1\leqslant i\leqslant t\tag{1}$$

式中　　$X_n(i-1)=[x(i-1),\ldots,x(i-n)]^T$　　(2)

RLS算法的递推公式为：

$$\hat{\varphi}(t)=\hat{\varphi}(t-1)+g_n(t-1)\eta_n(t)\qquad (t\geq 2)\tag{3}$$

式中，$\hat{\varphi}(t)$和$\hat{\varphi}(t-1)$分别是t时刻和t-1时刻的AR模型参数值，递推公式的初始值$\hat{\varphi}(1)=[\hat{\varphi}_1,\hat{\varphi}_2]^T$，(3)式中$\eta_n(t)$的为预报误差的暂时估值，

$$\forall t\in T;\tag{4}$$

$g_n(t-1)$为预报的增益向量，$g_n(t-1)=R_n^{-1}(t-1)X_n(t-1)$　　(5)

式中的　　$R_n(t-1)=\sum_{i=1}^{t-1}\lambda^{t-1-i}X_n(i)X_n^T(i)$　　(6)

上述公式中的$X(i)$为t时刻之前的网络流量，λ为减少旧数据影响的加权因子，即遗忘因子，经实验验证这里的$\lambda=0.9$最合适。

由于AR(2)模型就是2阶线性预报器，所以上述公式可以用于AR(2)模型参数的自动更新，从而模型具有自适应能力。

作者简介：崔伟兰(1980-)，女，长春工业大学硕士研究生，研究方向：计算机网络安全。尹逊伟(1980-)，男，中国防卫科技学院讲师。程永强(1978-)，男，长春工业大学硕士研究生。

1.3 异常检测

下面利用AR(2)模型来进行检测，AR(2)的模型是：

$$x_t=\varphi_1 x_{t-1}+\varphi_2 x_{t-2}+e_t$$

即 $e_t=x_t-\varphi_1 x_{t-1}-\varphi_2 x_{t-2}$

定义B是一步后移算子，即 $x_{t-1}=Bx_t$，则

$$e_t=x_t-\varphi_1 Bx_t-\varphi_2 Bx_{t-1}=x_t-\varphi_1 Bx_t-\varphi_2 B^2 x_t=(1-\varphi_1 B-\varphi_2 B^2)x_t=\varphi[B]x_t$$

其中，$\varphi[B]=1-\varphi_1 B-\varphi_2 B^2$

因此，$e_t=\varphi[B]x_t$，　t=1,2,…,N.

令 $\sigma^2=\frac{1}{N}\sum e_t^2$，并且定义 $e_{N+1}=\varphi[B]x_{N+1}$，　$\lambda=\frac{e_{N+1}}{\sigma}$

这里 σ^2 是时间序列中当前时刻向后N个相应的残差 e_t 平方和的平均值，λ表示当前观测值的残差与σ的比值，则σ作为检测是 x_{N+1} 否异常的统计量。当 $x_{N+1}<-L$ 或 $\lambda>U$ 时，x_{N+1} 就是异常值，这里的L和U是根据网络流量正常时的情况预先设定的大于零的常数。

实际检测过程中，当已经出现异常值时，必然会影响后面的检测，实验采取这样的处理方法：在下一次计算统计量λ时，该异常点的残差取前N个残差的平均值。尽管这样处理会损失一些信息，但却不会因这个异常值而影响下一次的检测。

2 算法分析

要实现实时检测，必须考虑算法的复杂性。第1步预处理过程在实际应用中可以离线完成，第2步、第3步和模型参数的自动更新以及异常检测需要在线完成。按N=20估算，零均值化有加减运算40次，主要的计算量在后3个阶段，第3步乘除运算约有88次，加减运算约有74次，模型参数的自动更新约有乘除运算20次，加减运算12次，异常检测阶段约有乘除运算64次，加减运算61次，全部在线运算总量约为乘除173次，加减运算187次，经测算，在Intel P IV 2.4G的机器上检测一次约需8μs，所以在现有的机器上进行实时检测是可行的。

3 实际网络环境下的流量异常检测

我们利用实验室的局域网对自适应AR模型进行实际测试，实验采用两个指标：检测率和虚警率。首先收集一周的正常流量数据，利用这些数据建立网络流量稳态模型。实验中采用TCP SYN洪水攻击，采集后两周内的跟踪数据，5分钟为一个时间间隔，两周中每天都在8：00-18：00这10个小时内的固定时刻人为插入攻击流量，每次攻击的持续时间为3分钟，攻击间的时间间隔为1个小时，这样在这个时间段内有10次攻击。为进一步研究不同攻击类型下的检测算法性能，本文分高强度攻击和低强度攻击两种情况进行测试。

第一次实验采用高强度攻击，其幅度是平均流量的2倍以上，用表1表示不同虚警率(False Rate)下的AR模型和自适应AR模型的检测结果。为更直观的看出检测效果，实验中引用ROC(Receiver Operating Characteristic)曲线，Y轴表示检测率DR(Detection Rate)，X轴表示虚警率FR(False Rate)。图1是高强度下的ROC曲线图，从图中可以更加清楚地看出自适应AR模型的在高强度攻击下的检测效果更好。

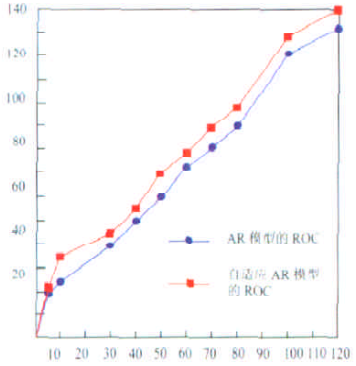

图1　高强度流量攻击下的检测率和虚警率 ROC 曲线

表1　高强度攻击时AR模型和自适应AR模型的攻击检测结果

FA	5	10	30	40	50	60	70	80	100	120
AR模型	19	25	39	50	60	72	81	90	121	132
自适应AR模型	22	29	44	55	69	78	89	98	128	140

第二次实验采用低强度攻击，强度为流量的0.5倍。低强度攻击检测相对来说要重要些，有两个原因：其一是对强度递增的攻击能够尽早检测到，其二是有助于在靠近攻击源的位置进行检测。检测结果和ROC曲线如表2和图2所示，可以看出在这种情况下自适应AR模型的检测效果更好。

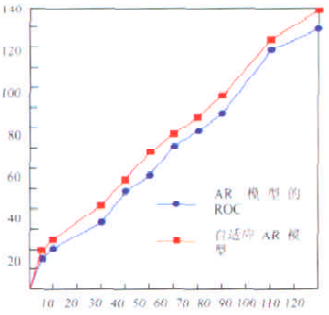

图2　低强度流量攻击下的检测率和虚警率 ROC 曲线

表2　低强度攻击时AR模型和自适应AR模型的攻击检测结果

FA	5	10	30	40	50	60	70	80	100	120
AR模型	15	20	33	48	56	70	78	86	118	128
自适应AR模型	19	24	41	54	67	75	85	95	123	138

综合图1和图2可看出，低强度流量攻击下，两种模型的检测能力都降低了，但自适应AR模型的检测能力只是略有降低，由此可看出自适应AR模型的检测性能更好。

4 结束语

本文提出了一种基于自适应AR模型的流量异常检测方法，该方法利用自适应滤波理论中的递归最小二乘算法，根据网络流量实时调整模型参数以及检测统计量，使模型更合理化。经实验验证，其算法复杂度低，能做到实时检测，并且因其能够对网络流量进行更准确的预测，所以能有效提高入侵检测的准确性，降低虚警率。

参考文献

[1]Simon Haykin著,郑宝玉等译.自适应滤波器原理[M].电子工业出版社.2003

[2]邹柏贤.一种网络异常实时检测方法[J].计算机学报.2003.