

# 大数据网络入侵过程的痕迹数据监测方法研究

张 凯

(重庆市气象局,重庆 401147)

**摘 要** 大数据网络数据规模巨大,对入侵过程痕迹数据进行监测的效率通常较低,一些带有入侵痕迹的数据特征在大数据环境下,特征逐渐淡化,当前方法无法在淡化的情况下准确采集痕迹数据的特点,无法形成待监测数据与痕迹数据之间的关系,导致监测效率和精度低下。提出一种基于模糊聚类概率的大数据网络入侵过程的痕迹数据监测方法,将采集的痕迹数据转换成频域信号,对其进行频谱或功率谱分析,依据时间变化的幅值将其转换成随频率变化的功率。采用核主元分析对痕迹数据信号特征进行提取,利用非线性转换将样本痕迹数据信号从输入空间映射至高维特征空间,在高维特征空间中通过 PCA 进行痕迹数据信号的频域特征提取。构建一个数学模型对特征模糊聚类概率进行描述,对待监测数据和痕迹数据之间的特征模糊聚类概率进行计算,通过衡量理论进行对比分析,使大数据网络入侵过程中的痕迹数据被完整的监测。实验结果表明,所提方法不仅所需时间少,而且监测精度高。

**关键词** 大数据网络 入侵过程 痕迹数据 监测  
**中图分类号** TP393.08; **文献标志码** A

网络入侵就是在未被授权的情况下,试图登录、处理或破坏网络系统的行为<sup>[1,2]</sup>。当大数据网络被入侵后,管理员不仅需进行安全加固和修复处理,还需对网络入侵痕迹数据进行监测,从而还原网络入侵过程及具体行为,获取证据线索与溯源,保留对网络入侵者依法追究责任的权力<sup>[3-5]</sup>。因此,对大数据网络入侵过程的痕迹数据进行监测具有重要意义,已经成为相关学者研究的重点课题,受到越来越广泛的关注<sup>[6]</sup>。

目前,有关大数据网络入侵过程痕迹数据监测的方法有很多,相关研究也取得了一定的成果,其中,文献[7]依据大数据入侵检测过程中痕迹数据的相关性,提出一种以局部监测为主的痕迹数据监测方法,对痕迹数据采样值的时间序列和事件随机过程特征进行采集,依据待监测数据与特征的符合程度对痕迹数据进行监测。该方法能够使节点之间的数据交换大大减少,然而其未对事件边缘节点进行监测;文献[8]提出一种基于直方图的大数据网络入侵过程的痕迹数据监测方法,该方法通过采集痕迹数据分布直方图信息,对冗余数据进行删除,对潜在的痕迹数据进行监测,大大降低了通信开销,但该方法未分析数据间的空间相关性,同时只适用于一维数据;文献[9]提出一种基于距离技术识别的大数据网络入侵过程的痕迹数据监测方法,该方法

采用聚集树结构以减少通信能耗,依据距离对痕迹数据进行识别,从而实现痕迹数据监测。然而因为该方法仅分析了痕迹数据间的时间相关性,不适用于大数据网络的痕迹数据监测;文献[10]提出一种基于聚类的大数据网络入侵过程的痕迹数据监测方法,该方法通过聚类将和痕迹数据相似的数据聚集在一起,无需数据分布的先验知识,但该方法无法为簇宽度参数设置有效的值,而且该方法主要依据数据实例间的距离计算,计算复杂度。

针对上述方法的弊端,提出一种基于模糊聚类概率的大数据网络入侵过程的痕迹数据监测方法,将采集的痕迹数据转换成频域信号,采用核主元分析对痕迹数据信号频域特征进行提取。对待监测数据和痕迹数据之间的特征模糊聚类概率进行计算,通过衡量理论进行对比分析使大数据网络入侵过程中的痕迹数据被完整的监测。实验结果表明,所提方法不仅所需时间少,而且监测精度高。

## 1 痕迹数据特征提取与检测原理分析

对大数据网络入侵过程的痕迹数据进行监测前,需对其特征进行提取,这里的特征主要指的是频域特征。将采集的痕迹数据转换成频域信号,对其进行频谱或功率谱分析,依据时间变化的幅值将其转换成随频率变化的功率。频谱的分析主要依据频率中心 $f_{FC}$ 、均方根频率 $f_{RMSF}$ 和根方差频率 $f_{RVF}$ ,分别代表了信号主频位置、主频变化和功率谱的集中程度,公式依次描述如下:

2016年1月28日收到

第一作者简介:张 凯(1981—),男,汉族,山东青岛人,工程师。研究方向:信息技术。E-mail: zkkxjsygc@163.com。

$$f_{FC} = \int_0^{+\infty} fS(f)df / \int_0^{+\infty} S(f)df \quad (1)$$

$$f_{RMSF} = \left[ \int_0^{+\infty} f^2 S(f)df / \int_0^{+\infty} S(f)df \right]^{1/2} \quad (2)$$

$$f_{RVF} = \left[ \int_0^{+\infty} (f - f_{FC})^2 S(f)df / \int_0^{+\infty} S(f)df \right]^{1/2} \quad (3)$$

式中,  $S(f)$  用于描述功率谱。则将采集的痕迹数据转换成频域信号的公式描述如下:

$$x_i = (f_{FC} + f_{RMSF} + f_{RVF})S(f) \quad (4)$$

所有痕迹数据信号的频域特征均可通过一个特征向量进行描述,由特征向量构成的空间被称作是特征空间。本节采用核主元分析(KPCA)对痕迹数据信号特征进行提取,该方法为主元分析的非线性推广,其基本思想如下:利用非线性转换将大数据网络入侵过程的样本痕迹数据信号从输入空间映射至高维特征空间,再在高维特征空间中通过 PCA 进行痕迹数据信号的频域特征提取,下面进行详细的分析。

假设  $x$  为  $n$  维痕迹数据向量,  $\{x_i, i=1, 2, \dots, N\}$  用于描述  $x$  的一个痕迹数据信号样本集,通过非线性  $H$  将样本痕迹数据信号从空间  $R^n$  映射至高维特征空间  $R^l$ ,再在该高维特征空间中实现主成分分析。

假设  $H(x_i)$  已去均值  $\sum_{i=1}^N H(x_i) = 0$ , 则  $H(x_i)$  的协方差矩阵  $C$  可描述成

$$C = \frac{1}{N} \sum_{i=1}^N H(x_i) H(x_i)^T \quad (5)$$

式(5)中的特征值与特征向量之间的关系可描述成

$$\lambda_k v_k = C v_k \quad (6)$$

式(6)中,特征值  $\lambda_k \geq 0$ ,  $v_k (k=1, 2, \dots, t)$  用于描述特征向量。

将式(6)代入式(5)有

$$C v_k = \frac{1}{N} \sum_{i=1}^N H(x_i) \langle H(x_i), v_k \rangle = \lambda_k v_k \quad (7)$$

式(7)中,内积  $\langle H(x_i), v_k \rangle = H(x_i)^T v_k$ 。

令和全部非零特征值  $\lambda_k$  相应的特征向量  $v_k$  处于  $\{H(x_i), i=1, 2, \dots, N\}$  形成的平面中,以存在不全为 0 系数  $[T_i, i=1, 2, \dots, N]$ , 使

$$v_k = \sum_{i=1}^N T_i H(x_i) \quad (8)$$

通过式(6)~式(8)可获取

$$\begin{aligned} \lambda_k \langle H(x_i), v_k \rangle &= \lambda_k \sum_{j=1}^N T_j \langle H(x_i), H(x_j) \rangle = \\ &= \langle H(x_i), C v_k \rangle = \\ &= \frac{1}{N} \sum_{s=1}^N \{ \langle H(x_s), H(x_i) \rangle \sum_{j=1}^N T_j \times \end{aligned}$$

$$\langle H(x_s), H(x_j) \rangle \} \quad (9)$$

假设  $N \times N$  矩阵用  $K_{ij} = k(x_i, x_j) = \langle H(x_i), H(x_j) \rangle$  进行描述,其中  $k(x_i, x_j)$  为符合 Mercer 定理的核函数,将式(5)化简成  $N \lambda_k K T = K^2 T$ , 则有

$$N \lambda_k T = K T \quad (10)$$

式(10)中,  $T = [T_1, T_2, \dots, T_N]^T$ 。

则  $K$  的特征值与特征向量可描述成  $N \lambda_k$  与  $T^k$ ,  $k=1, 2, \dots, N$ 。将特征值按照从大到小的顺序排列,如果前  $m$  个特征值之和占总特征值和的比值超过既定阈值,则认为主元数量是  $m$ 。

为了对特征向量  $v_k$  进行归一化处理,还需对  $T$  进行规范化:  $\bar{T}^k = T^k / \sqrt{\lambda_k}$ , 则可获取测试痕迹数据信号样本  $x_i$  在  $R^m$  空间中的第  $k$  个主向量  $v_k$  上的投影,也就是  $x_i$  的特征值

$$v_k = \sum_{i=1}^N \bar{T}_i^k K(x, x_i) \quad (11)$$

通过以上方法完成数据特征的提取与检测。

## 2 痕迹数据监测实现过程分析

在对大数据网络入侵过程的痕迹数据进行监测的过程中,最关键的步骤为求出待监测数据与痕迹数据之间的特征模糊聚类概率,需在上节提取的痕迹数据特征值的基础上,构建一个数学模型对特征模糊聚类概率进行描述,该模型主要由  $n$  个待监测数据与  $p$  条相关性较大的痕迹数据构成,塑造一个大小为  $n \times p$  的矩阵,用  $C_{n \times p} = \{c_{jk}\}_{n \times p}$  进行描述。

$$C_{n \times p} = \begin{bmatrix} c_{11} & c_{12} & \cdots & c_{1p} \\ c_{21} & c_{22} & \cdots & c_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ c_{n1} & c_{n2} & \cdots & c_{np} \end{bmatrix} \quad (12)$$

为了得到大数据网络入侵过程的待监测数据和痕迹数据之间的关系,简化计算过程,减少运算时间,需要对矩阵进行降维处理

$$C_l = (w_j + \rho_j) / l \quad (13)$$

式(13)中,  $\rho_j$  用于描述矩阵相关性系数;  $w_j$  用于描述权值系数。若在符合上述条件的基础上,  $l$  无限接近于  $\text{rank}(B)$ , 则  $C_l$  和  $C_{n \times p}$  将无限接近。完成对矩阵的降维处理后,则可获取与监测痕迹数据相关的节点数据,过滤掉很多无关数据,提高了痕迹数据监测效率。

假设  $p$  为大数据网络入侵过程中痕迹数据的数量;  $e_k$  为第  $k$  个扫描节点。按照相关性将痕迹数据划分成  $L$  类,其特征为  $(v_1, v_2, \dots, v_n)$ ,  $n$  用于描述待监测数据的数量,  $e_k (v_{k1}, v_{k2}, \dots, v_{kn})$  用于描述网络数据,则  $e_k$  为大数据网络入侵过程中所有待监测数

据。依据特征模糊聚类概率对痕迹数据进行监测,需将待监测数据  $e_k$  依据不同的特征进行分类,判断大数据网络中的数据是否为痕迹数据时,需通过式(14)对待监测数据和痕迹数据之间的特征模糊聚类概率进行计算

$$Q(f_k) = q(f_k)q(f_i + \chi_n)/q(c_j) \quad (14)$$

式(14)中,  $q(c_j)$  用于描述对不同种类待监测数据进行监测时,获取的数据和痕迹数据之间的特征匹配概率;  $q(f_k)$  用于描述待监测数据与入侵数据的匹配概率;  $\chi_n$  为常量,用于调节概率参数。若监测数据相同,则  $q(f_k)$  将保持不变;  $q(f_k c_j)$  用于描述待监测数据和痕迹数据之间的关系。上述变量的公式描述如下:

$$q(f_k c_j) = q(w_1 c_j) + q(w_2 c_j) + \cdots + (w_n c_j) \quad (15)$$

$$q(c_j) = q(f_k)q(c_j f_k)/n \quad (16)$$

$$q(f_k) = T(p(c_j)f_k)/T_{ek} \quad (17)$$

式中,  $T(p(c_j)f_k)$  用于描述  $e_k$  类中出现的痕迹数据的数量;  $T_{ek}$  用于描述大数据网络入侵过程中待监测集合中属于痕迹数据的数量。  $j$  用于描述待监测样本数量。将  $Q(f_k)$  超过既定阈值的待监测数据看作是痕迹数据,从而实现大数据网络入侵过程中痕迹数据的监测。

为了使大数据网络入侵过程中的痕迹数据被完整的监测,需通过衡量理论进行对比分析,使其满足下述条件

$$q(f = \sin\beta) < a \quad (18)$$

式(18)中,  $a$  用于描述痕迹数据衡量标准。在大数据网络入侵过程的痕迹数据进行监测时,痕迹数据与其他数据相关性概率之和为 1,也就是

$$q(f = \sin\beta) + q(f = \cos\beta) = 1 \quad (19)$$

依据(19)式有

$$q(f = \cos\chi) < a/(a-1) \quad (20)$$

完成对  $a$  的优化处理后,若  $q(f = \cos\chi)$  无限接近于 1,则痕迹数据监测效果将较好。

### 3 实验结果仿真

为了验证本文提出的基于特征模糊聚类概率的大数据网络入侵过程痕迹数据监测方法的有效性,需要进行相关的实验分析。实验将马尔科夫方法作为对比进行分析。在 Window 7.0, CPU 是 Intel Pentium Dual Core 的微机环境中,通过 Oracle9. 2 数据库管理系统进行编程,完成下述模拟实验。

分别采用本文方法和马尔科夫方法对大数据网络入侵过程的痕迹数据进行监测,对两种方法的查全率和查准率进行比较,得到的结果用表 1 进行描述。

表 1 两种方法监测性能比较

Table 1 Two methods of monitoring the performance comparison

| 数据集数量  | 本文方法  |       | 马尔科夫方法 |       |
|--------|-------|-------|--------|-------|
|        | 查全率/% | 查准率/% | 查全率/%  | 查准率/% |
| 5 000  | 98    | 99    | 82     | 79    |
| 10 000 | 92    | 97    | 76     | 76    |
| 15 000 | 95    | 98    | 71     | 81    |
| 20 000 | 97    | 96    | 73     | 85    |
| 25 000 | 93    | 97    | 84     | 87    |
| 30 000 | 91    | 98    | 68     | 82    |
| 35 000 | 95    | 99    | 70     | 77    |
| 40 000 | 94    | 97    | 65     | 74    |

分析表 1 可以看出,采用本文方法对大数据网络入侵过程的痕迹数据进行监测,得到结果的查全率与查准率均远远高于马尔科夫方法,说明本文方法的监测精度较高,验证了本文方法的有效性。

分别采用本文方法和马尔科夫法对大数据网络入侵过程的痕迹数据进行监测,将得到的结果和实际结果进行比较,给出监测误差曲线,用图 1 进行描述。

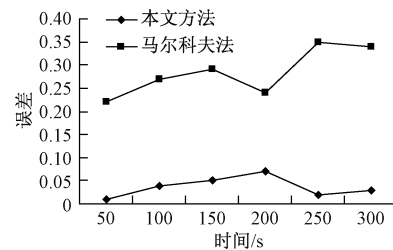


图 1 两种方法监测误差曲线比较结果

Fig. 1 Two methods of monitoring error curve comparison results

分析图 1 可以看出,采用本文方法对大数据网络入侵过程的痕迹数据进行监测,得到的误差曲线在 0.05 上下波动,而采用马尔科夫法得到的误差曲线基本在 0.2 以上,明显高于本文方法,说明本文方法对痕迹数据的监测精度较高。

为了进一步评价两种方法的性能,采用 ROC (receiver operating characteristic) 曲线对监测正确率与误报率之间的关系进行描述,两种方法的 ROC 曲线用图 2 进行描述,横坐标代表误报率,纵坐标代表准确率。

分析图 2 可以看出,采用本文方法得到的 ROC 曲线监测准确率达到 90% 以上时,马尔科夫法的 ROC 曲线监测准确率只有 70% 左右,同时本文方法的误检率明显低于马尔科夫法,进一步验证了本文方法的监测性能。

分别采用本文方法和马尔科夫法对本文方法和

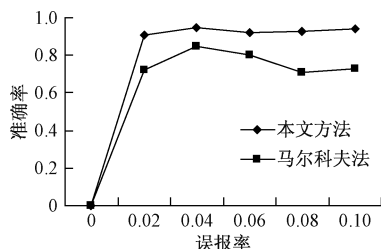


图2 两种方法 ROC 曲线比较结果

Fig.2 ROC curve comparison results of the two methods

马尔科夫法进行相同实验所需的时间进行比较,得到的结果用表2 进行描述。

表2 两种方法效率比较结果

Table 2 Two methods of efficiency comparison results

|   | 本文方法/s | 马尔科夫法/s |
|---|--------|---------|
| 1 | 2.5    | 4.9     |
| 2 | 3.1    | 7.2     |
| 3 | 2.7    | 6.5     |
| 4 | 4.2    | 8.1     |
| 5 | 5.9    | 9.5     |
| 6 | 5.1    | 8.7     |
| 7 | 6.2    | 11.4    |
| 8 | 4.8    | 7.9     |

分析表2 可以看出,在相同的实验条件下,采用本文方法对大数据网络入侵的痕迹数据进行监测所需的时间明显少于马尔科夫法,这是因为本文方法分析了正常数据与痕迹数据之间的关系,大大减少了计算量。

## 4 结论

本文提出一种基于模糊聚类概率的大数据网络入侵过程的痕迹数据监测方法,将采集的痕迹数据转换成频域信号,对其进行频谱或功率谱分析,依据时间变化的幅值将其转换成随频率变化的功率。采用核主元分析对痕迹数据信号特征进行提取,利用非线性转换将样本痕迹数据信号从输入空间映射至高维特征空间,在高维特征空间中通过 PCA 进行痕迹数据信号的频域特征提取。构建一个数学模型对特征模糊聚类概率进行描述,对待监测数据和痕迹数据之间的特征模糊聚类概率进行计算,通过衡量理论进行对比分析使大数据网络入侵过程中的痕迹数据被完整的监测。实验结果表明,所提方法不仅所需时间少,而且监测精度高。

## 参 考 文 献

- 王曙霞. 大数据环境下的网络主动入侵检测方法研究. 科技通报, 2015; 31(8): 225—227  
Wang S X. Big data under the environment of network intrusion detection method actively study. Science and Technology, 2015; 31(8): 225—227
- 宋文超, 王 焯, 黄 勇, 等. 大数据环境下的云计算网络安全入侵检测模型仿真. 中国西部科技, 2015; (8): 86—88  
Song W C, Wang Ye, Huang Yong, et al. Under the environment of big data cloud computing network intrusion detection model simulation. Journal of Safety Science and Technology in Western China, 2015; (8): 86—88
- 曹 旭, 曹瑞彤. 基于大数据分析的网络异常检测方法. 电信科学, 2014; 30(6): 152—156  
Cao Xu, Cao R T. Network anomaly detection method based on large data analysis. Journal of Telecom Science, 2014; 30(6): 152—156
- 陈玉明, 谢斐星, 吴克寿, 等. 基于邻域关系的网络入侵检测特征选择. 常州大学学报: 自然科学版, 2014; 26(03): 1—5  
Chen Y M, Xie F X, Wu K S, et al. The network intrusion detection based on neighborhood relation feature selection. Journal of Changzhou University: Natural Science Edition, 2014; 26(3): 1—5
- 李天枫, 姚 欣, 王劲松. 大规模网络异常流量实时云监测平台研究. 信息安全, 2014; (9): 1—5  
Li T F, Yao Xin, Wang J S. Large scale network anomaly traffic real-time monitoring platform of cloud study. Information Network Security, 2014; (9): 1—5
- 程 源, 楚春颖. 一种基于 PSO 辨别树的 P2P 网络入侵检测方法. 科技通报, 2013; (6): 56—58  
Cheng Yuan, Chu C Y. A P2P network intrusion detection method based on PSO to identify trees. Science and Technology, 2013; (6): 56—58
- 阳时来, 杨雅辉, 沈晴霓, 等. 一种基于半监督 GHSOM 的入侵检测方法. 计算机研究与发展, 2013; 50(11): 2375—2382  
Yang S L, Yang Y H, Shen Q N, et al. An intrusion detection method based on a semi-supervised GHSOM. Journal of Computer Research and Development, 2013; 50(11): 2375—2382
- 袁遇晴, 况湘玲, 凌利军. 基于数据挖掘的网络入侵检测研究. 计算机安全, 2014; (7): 14—17  
Yuan Y Q, Kuang X L, Ling L J. Network intrusion detection based on data mining research. Computer Security, 2014; (7): 14—17
- 吴 琼. 云计算环境下的联合网络入侵检测方法仿真. 计算机仿真, 2015; 32(6): 276—279  
Wu Qiong. Cloud computing environment of network intrusion detection method Simulation. Computer simulation, 2015; 32(6): 276—279
- 王登贵. 基于 MCU 的水果贮藏室温湿度监测及报警系统设计. 电子设计工程, 2015; (19): 14—17  
Wang D G. Fruit storage temperature and humidity monitoring and alarm system based on MCU design. Journal of Electronic Design Engineering, 2015; (19): 14—17

## Big Data Network Intrusion Traces of Process Data Monitoring Method Research

ZHANG Kai

(Chongqing Meteorological Bureau, Chongqing 401147, P. R. China)

[**Abstract**] Big data network data size, traces the process of intrusion data monitoring efficiency is low, often some data with invasion of trace characteristics under the environment of big data, characteristic gradually fade out, under the condition of current method can't played down the characteristics of accurate gathering trace data, unable to form for monitoring data and trace data, the relationship between the monitoring efficiency and low accuracy. A kind of big data network was put forward based on fuzzy clustering probability of process data monitoring method, the trace of the collected data into frequency domain signal, the spectrum and power spectrum analysis, according to the time change amplitude convert them to change with frequency power. Using Kernel principal component analysis to trace data signal characteristic extraction, using nonlinear transformation to trace sample data signals from the input space is mapped to high-dimensional feature space, in the high dimensional feature space by PCA to trace data signal in the frequency domain feature extraction. Build a mathematical model to simulate the characteristics of fuzzy clustering probability description, treatment of monitoring data and trace data between the characteristics of the fuzzy clustering probability calculation, by comparing the measure theory makes big data in the process of network intrusion trace monitoring data is complete. The experimental results show that the proposed method is not only less time required, and monitoring of high precision.

[**Key words**] big data network the invasion process trace data monitoring

(上接第 253 页)

- 15 马亲民,王晓春,戴光智.无线传感器网络面临的攻击与对策.传感器与微系统,2012;31(3):8—10,14
- Ma Q M, Wang X C, Dai G Z. Attacks and countermeasures faced by WSNs. Transducer and Microsystem Technologies, 2012;31(3):8—10,14
- 16 蒋云霞,符琦.无线传感器网络(WSNs)路由安全问题的现状与对策研究.中国安全科学学报,2008;18(12):117—123
- Jiang Y X, Fu Q. Current status of routing security of WSNs and its

countermeasures. China Safety Science Journal, 2008;18(12):117—123

- 17 李成法,陈贵海,叶懋,等.一种基于非均匀分簇的无线传感器网络路由协议.计算机学报,2007;30(1):27—36
- Li C F, Chen G H, Ye M, et al. An uneven cluster-Based routing protocol for wireless sensor networks. Chinese Journal of Computers, 2007;30(1):27—36

## A Security Communication Strategy for WSNs with both Active and Passive Defenses

SUN Jia-wen, YANG Bo, JIA Xin-chun

(School of Mathematical Sciences, Shanxi University, Taiyuan 030006, P. R. China)

[**Abstract**] Because of its nature features such as openly communication channel, resource-constrained nodes and randomly deployment, the problem of security is facing challenge for wireless sensor networks. In this paper, a security communication strategy which possesses both active and passive defenses is proposed, for the problem of some applications of WSNs which are always placed in dangerous environment that requires high safety. By combining the technology of active defenses which adopt encryption, decryption technique, signature certification, integrity identify, and with passive defenses of transforming the base-station to create a relatively safe operate environment for network. The performance analysis shows this strategy can reduce the possibility of the network under attack, decrease the effect of attack, improve the network anti-dilapidated ability and prolong the lifetime.

[**Key words**] wireless sensor networks(WSNs) security active defense passive defense