

引用格式:张杰,傅文博. 网络安全数据可视化融合的分析方法[J]. 微电子学与计算机,2019,36(6):101-104. [ZHANG J,FU W B. Analysis method of visual fusion of network security data [J]. Microeletronics & Computer,2019,36(6): 101-104.]

网络安全数据可视化融合的分析方法

张 杰,傅文博

(山西大同大学 计算机与网络工程学院,山西 大同 037009)

摘 要: 文章提出一种基于时间序列的网络安全数据可视化融合分析方法,利用网络数据之间的关联性进行特征选择和一致性过滤,保留原始特征的同时消除冗余特征;通过网络数据特征之间的信息熵计算,利用 INTERACT 方法实现归一化处理;通过时间序列分析法对处理后的网络数据进行可视化融合分析,利用切比雪夫法对网络数据包进行验证,判断是否存在入侵行为,实现网络安全数据的可视化融合. 实验结果表明,通过对检测阈值的调控能够提高检测率,随着数据规模的不断扩大,所提方法的检测率较为平稳,且误报率相比其他方法更低.

关键词: 网络;安全数据;可视化融合;分析

中图分类号: TP393

文献标识码: A

文章编号: 1000-7180(2019)06-0101-04

Analysis method of visual fusion of network security data

ZHANG Jie,FU Wen-bo

(School of Computer and Network Engineering,Shanxi Datong University,Datong 037009,China)

Abstract: In this paper a visual fusion analysis method for network security data based on time series is proposed. The correlation between network data is used for feature selection and consistency filtering. The original features are preserved while the redundant features are eliminated. The information entropy between network data features is proposed. Calculation, using the INTERACT method to achieve normalization processing; through the time series analysis method for visual fusion analysis of the processed network data, using the Chebyshev method to verify the network packets to determine whether there is intrusion behavior, to achieve network security data Visual fusion. The experimental results show that the detection rate can be improved by regulating the detection threshold. With the continuous expansion of data size, the detection rate of the proposed method is relatively stable, and the false alarm rate is lower than other methods.

Key words: The internet;safety data;visual fusion;analysis

1 引言

网络应用的迅速发展对网络安全提出了更高的要求^[1]. 随着人们对网络安全需求的不断增加,在分析网络安全数据时网络安全分析人员遇到了许多新的难题. 文献[2]提出一种基于网络异常检测的分析方法,利用转移学习方法对熟知的网络进行实验检验,来验证网络中是否存在异常入侵行为进行数据

可视化的分析. 但是这些检测方法有两种缺点,一是检测率低,二是误报率高.

根据上述方法所存在的缺点,提出了基于时间序列的网络安全数据可视化融合分析方法.

2 网络安全数据可视化的研究和分析方法

2.1 网络安全数据的预处理

由于筛选方法挑选的特征只取决于统计数据培

收稿日期: 2018-10-24; 修回日期: 2018-11-17

基金项目: 山西省教育科学“十三五”规划项目(GH-18045);大同市工业重点研发计划项目(2017011);山西大同大学科研基金项目(2017K12)

训结果,比较而言成本预估低的算法让它对于类似 KDDCup99 的整个数据包有较高效率^[3,4].

(1) 基于启发式评估函数的网络数据特征提取

利用关联的启发式评估函数对网络数据进行特征选择,提取出与分类紧密相关且网络数据之间不具有强关联性的单个特征组成最优特征子集,消除网络数据的冗余特征,实现网络数据的有效特征提取.具体描述公式如下:

$$M_s = \frac{k_{r_{ff}}}{\sqrt{k+k(k-1)}r_{ff}} \quad (1)$$

式中, $r_{r_{ff}}$ 和 r_{ff} 分别表示网络数据特征与所属类别和网络数据特征与联系强弱的平均值; M_s 代表一个数据包包含 k 个网络特征子集的引导式的重要性; $k_{r_{ff}}$ 代表网络数据特征与类别接触反映出来的强弱值; $\sqrt{k+k(k-1)}r_{ff}$ 代表在特征处理中有多少多余的网路数据.

(2) 原始网络数据的信息熵计算

原始网络数据是基于 SU 方法的数据信息相对不确定性.此性质是讲述 x, y 中原始信息的利润 IG 和熵 H 比值的度量.计算公式得出:

$$SU(x, y) = 2 \frac{IG(\frac{x}{y})}{H(x) + H(y)} - k_{r_{ff}} \quad (2)$$

$$IG(\frac{x}{y}) = H(y) + H(X) - H(x, y) \quad (3)$$

除了 SU, INTERACT 还包含原始属性一致性.网络特征属性的一致性能反映出此项网络特征对一致性的影响达到的标准^[5].此方法通常在查找数据特征中使用,且彼此之间互相影响.由两部分组成:一是上述网络特征都以不确定性的降序排列,二是从后面启动网络数据特征序列,分别评估网络数据特征,假设阈值高于网络数据特征的一致性,则阈值就不会被使用,相反阈值被使用,得到去除冗余数据的特征,公式如下:

$$X' = IG(\frac{x}{y}) - M_s \quad (4)$$

2.2 基于时间序列的网络数据可视化融合分析

网络数据的时间序列是监管检测过程中按照时间先后顺序得到的.用 $x(i)$ 表示监管检测数据,则 $x(1), \dots, x(i), \dots$ 表示每个时间段内 K 次检测的时间序列定为一个周期,划分出每个周期运行的时间段, $x(i, j)$ 代表检测时第 j 个时段内入侵行为发生的概率. $M(j)$ 表示在第 j 个时段内这个时间发生的平均次数,则发生平均次数的表达式为:

$$M(j) = \frac{1}{K} \sum_{i=1}^K X(i, j) + y \quad (5)$$

$K+1$ 次检测之后,用 $M(j)$ 表示第 j 个时间段内事件(数据包的接收率或数据包到达的时间间隔)的预期值.用 $x(K+1, j)$ 代表第 j 个时刻真实所发生事件的次数,则得到的表达式如下:

$$z(j) = \frac{|x(K+1, j - M(j))|}{M(j)} \quad (6)$$

根据时间序列的网络数据建立时间模型,设置一个数据的阈值用 δ 表示,当阈值偏差 $z(j)$ 大于且等于 δ_1 时,即可得知入侵事件大概会在什么时间段发生入侵.

通过多项式对时间序列划分出可用的阈值,对入侵事件分析预测.在时间序列网络安全数据 Y 中取值, D 代表领域布置, A 代表节点收集, m 代表网络数据长度 d 代表数据的维度, W 代表数据窗口, ω 代表流动窗口 ($\omega \ll m$)^[6].检测时间序列的网络数据可视化过程表达式为:

$$c_0 = \frac{1}{\omega} \sum_{j=1}^m f(t_j) p_0(t_j) = \frac{1}{\omega} \sum_{j=1}^m f(t_j) Z_j \quad (7)$$

$$c_i = \frac{2}{\omega} \sum_{j=1}^m f(t_j) p_i(t_j) \quad (8)$$

类似序列: S, R 的时间序列数据分别为 C, D .假设 $(C, D) < \epsilon$, 则称时间序列 S 和 R 相似.对时间序列 S 来说,假如历史数据的窗口中,与它类似的时间序列 $similar(S, R)$ 表示数量小于一定的比例,那么时间序列 S 称为离群时间序列,即:

$$\frac{|\{R | similar(S, R), S, R \in W\}| + \sum_{j=1}^m f(t_j)}{|\{H | H \in W\}|} \quad (9)$$

根据网络数据特征的预处理结果对网络数据进行熵计算,时间序列内数据预处理到达的时间和网络数据接收成功的概率称为其特征,从而为网络入侵的判断打好基础.网络数据的时间序列方法可被网络入侵检测过程使用^[7,8].则其方法如下:

对某个时间段的限流开始检测; $z(j)$ 设为 j 时间的差值; t 设为无异常时时间序列的个数, δ_1 设为阈值的时间序列检测, S 设为时间序列在 j 时间内的数据运行结果, δ_2 设为阈值与时间序列的近似值.

在规定范围内检测到网络数据集合,警惕对网络的入侵.选择各种时刻所对应的时间序列,设 δ_1 为阈值,设 $z(j)$ 为时间差值,根据以下公式进行网络安全数据可视化融合分析:

$$z(j) < \delta_1 \quad (10)$$

$$z(j) \geq \delta_1 \quad (11)$$

$S \geq \delta_1$ (12)

$S \geq \delta_2$ (13)

如果满足上述公式(10),则该测量时间序列在
规定范围内发送给正常的行为库,从而对正常数据
进行放行;如果满足上式(11),则说明 j 时间内可能
存在入侵,时间序列内的网络数据可能存在偏差,通
过分析来判断 j 时间内的时间序列 S 和阈值 δ_2 . 式
(12) 成功验证,开始分析相对偏差,式(13) 成功验
证,则立刻将时间序列 S 设为离散时间序列,从而实
现数据的入侵行为检测,报警系统启动.

3 实验及结果分析

3.1 实验环境与设计

在 Corei3 PC 机(2G 内存)的 WindowsXP 环境下,
利用网络测试基地,对主要内容进行分析. 收集一组网
络数据,依次选择不同的网络数据进行实验,数据规模
可以设置为 10 组、50 组、80 组、120 组、200 组.

把误报率和检测率作为主要衡量标准,开始验
证误报率和检测率的变化趋势并对检测方法进行评
价,检测率代表样本数与样本总数之比,误报率代表
被误判的正常样本数与正常样本总数之比.

3.2 检测阈值对不同时间序列分析方法功能的
影响

针对检测阈对不同时间序列分析的基本特征,
设定几个数据大小不同的参数,得出 $\delta_1 \in$
[1.2,5.2]的选值开始实验,对参数 δ_1 的取值分析
并探讨,分析 δ_1 是否会影响时间序列分析方法,检
验结果可以看出,随着 δ_1 取值的不停增大,时间序
列分析方法的检测率适当降低,当 $\delta_1 > 4$ 时,检测率
快速下降,而当 $\delta_1 < 3$ 时,它的误报率偏高.

为了更详细地比较参数 δ_1 对时间序列分析方
法性能的影响,我们主要对上述五种不同模式的数据
集进行实验. 在实验中得知,取值的大小影响样本
集的变化,同时检测率、误报率也产生了应有的变
化. 实验结果表明,数据集用 $S_1 S_2 S_3 S_4$ 及 S_5 表示,本
次实验主要选值 δ_1 分别取 2、2.5、3、3.5 这四个值开
始验证. 如表 1、表 2 所示.

表 1 不同阈值所对应的方法的检测率(%)

阈值 δ 测试集	检测率(%)				
	S1	S2	S3	S4	S5
2	98	98	97	98	98
2.5	92	92	88	90	90
3	89	87	89	88	88
3.5	60	65	65	64	65

表 2 不同阈值所对应的方法的误报率(%)

阈值 δ 测试集	检测率(%)				
	S1	S2	S3	S4	S5
2	1.5	1.6	1.6	1.5	1.6
2.5	1	0.9	1	0.9	0.9
3	0.8	0.6	0.6	0.8	0.8
3.5	0.5	0.4	0.5	0.3	0.4

从表里很容易看出,阈值 δ_1 取 2 时,它的检测
率比较高,平均达到 98%,但误报率相对也高,平均
达到 1.6%. 当阈值取 3.5 时,进行异常判断的区域
变小,区域内存在很多异常点不容易通过此方法辨
识出来,则检测率比较低,误报率很低下滑到 0.5%
以下,更加能得出大于 3 方法性能不完美,分析结果
可得到,取值在 2.5 和 3 时,能保证此方法在维持高
检测率同时也最大化降低了误报率.

3.3 数据集的规模对方法性能的影响

根据上面实验的参数 δ_1 对时间序列分析方法
性能的影响,取阈值为 3 当做参考,比较文献[2]和
时间序列分析两种方法对 5 个不同规模的实验样本
进行多次实验.

图 1 表示误报率的对比. 横坐标分别用 S_1 — S_5
来表示,纵坐标对照的是检测率和误报率,黑色矩形
代表文献[2]方法,白色则代表时间序列分析方法.

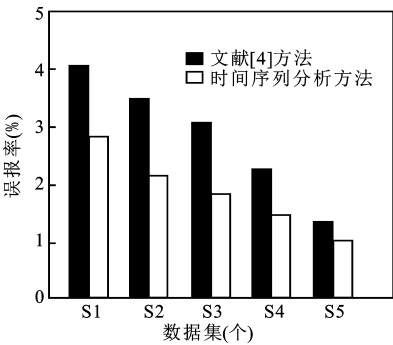


图 1 两种方法误报率的对比

随着数据集规模的不断增大,时间序列分析方
法的检测率逐渐升高,误报率相对减少,由于样本数
的增加,正常与异常数据点具有显著的区分. 这时,
不正常区域范围里的数据点比正常的数据点的数量
少了很多,由此说明文献[2]方法可以很轻松辨识大
多数的异常值. 但时间序列分析方法的检测率一直
都比文献[2]方法的检测率高,还有误报率比文献
[2]方法低. 是由于时间序列分析方法在文献[2]方
法的基础上建立的,通过阈值的调整,让这个方
法能更精准判断,可以辨识到文献[4]方法不能辨识的

异常值.

4 结束语

提出了基于时间序列分析的网络安全数据可视化融合方法,简单说明了分析人员对网络数据的应对,并主要从选择特征和网络安全数据可视化这两个方面进行验证,利用时间序列分析方法来判断网络情况,对检测率和误报率进行了实验,实验中能够根据阈值的变化完成调整,提升了时间序列分析方法的检测率,并使误报率降低.

参考文献:

[1] 胡晓宏. 基于图像处理的设备异常检测关键技术仿真[J]. 计算机仿真, 2016, 33(1):425-429.

[2] 赵新杰, 刘渊, 孙剑, 等. 基于迁移学习和 D-S 理论的网络异常检测[J]. 计算机应用研究, 2016, 33(4): 1137-1140.

[3] 胡海峰, 周改云, 刘云霞. 基于过往入侵特征分析潜在网络威胁检测仿真[J]. 计算机仿真, 2016, 33(2): 322-325.

[4] 邱卫, 杨英杰, 汪永伟, 等. 基于改进遗传算法和隐 Markov 模型的协议异常检测方法[J]. 计算机应用研究, 2016, 33(4):1164-1168.

[5] 庄翠芳, 阳波, 牛俊泽. 多频率激励生物电阻抗测量方法的研究[J]. 计算机测量与控制, 2016, 24(5):71-73.

[6] 王德文, 杨力平. 智能电网大数据流式处理方法与状态监测异常检测[J]. 电力系统自动化, 2016, 40(14):122-128.

[7] 肖云, 何金海, 许爱华, 等. 江西省三类强对流天气环境物理量对比分析[J]. 科学技术与工程, 2016, 16(14):107-114.

[8] 黄海军. 基于云计算的网络安全评估[J]. 电子设计工程, 2016, 24(12):115-117.

作者简介:

张 杰 男, (1979-), 硕士研究生, 讲师. 研究方向为网络安全、物联网. E-mail:ss356398632@163.com

傅文博 男, (1976-), 硕士, 副教授. 研究方向为网络安全、物联网.