网络出版时间：2017-07-31 18:50:38
网络出版地址：http://kns.cnki.net/kcms/detail/32.1397.N.20170731.1850.004.html

基于大数据流的网络流量检测与分析

程伟华[1],赵 军[2],吴 鹏[1]

(1. 江苏电力信息技术有限公司,江苏 南京 210024;2. 国网江苏省电力公司,江苏 南京 210029)

摘 要:针对网络流量异常检测问题,该文提出1种新的网络流量检测和分析系统。采用分布式流式处理机制达到实时检测。利用大数据平台分布式存储、数据计算分析的能力,实现网络数据分布式存储,训练网络数据协议特征库。在江苏省电力公司的营销、运行与调度等业务场景中,该网络流量检测与分析系统取得了很好的实际效果,为各个业务场景的分析提供了业务支撑。

关键词:数据包分析;异常检测;大数据流;网络流量检测;分布式流式处理机制;大数据平台;分布式存储

中图分类号:TP393.08 **文章编号**:1005-9830(2017)03-0294-07

DOI:10.14177/j.cnki.32-1397n.2017.41.03.004

Network traffic detection and analysis based on big data flow

Cheng Weihua[1], Zhao Jun[2], Wu Peng[1]

(1. Jiangsu Electric Power Information Technology Co., Ltd., Nanjing 210024, China;
2. State Grid Jiangsu Electric Power Company, Nanjing 210029, China)

Abstract: A new network traffic detection and analysis system is proposed for network traffic anomaly detection problem. A distributed stream processing mechanism is used to achieve a real-time detection ability. Network data distributed storage is achieved and a network protocol feature library is trained by using the distributed storage and the data computational analysis ability of a big data platform. The network system of detection and analysis gains a good performance in the business of marketing, operation and dispatching in Jiangsu Electric Power Company, and provides a good support for the analysis of various business scenarios.

Key words: data packet analysis; anomaly detection; big data flow; network traffic detection; distributed stream processing mechanism; big data platform; distributed storage

收稿日期:2016-11-22 **修回日期**:2016-12-30

基金项目:国网江苏省电力公司科技项目(SGJSXTOOYJYJ1588925)

作者简介:程伟华(1978-),男,高级工程师,主要研究方向:计算机软件与理论、电力信息化,E-mail:chengweihua78@126.com。

引文格式:程伟华,赵军,吴鹏. 基于大数据流的网络流量检测与分析[J]. 南京理工大学学报,2017,41(3):294-300.

投稿网址:http://zrxuebao.njust.edu.cn

随着 Internet 技术的发展,网络规模不断扩大,应用领域不断扩展,由此产生的各种安全问题日益凸显,安全问题是互联网技术领域最受关注的问题之一。这些问题不仅来自于外部网络,如计算机病毒、黑客以及网络陷阱等;也有很大一部分来自于网络内部,如应用系统错误、操作系统的安全漏洞和缺陷或者数据库自身存在的隐患等。在应对网络威胁时,传统的入侵检测系统通常只针对外部网络中的病毒、黑客等。该系统在预防攻击时需先掌握外部攻击的特征,并需要经常更新数据库,这种做法已经越来越无法适应高速网络的要求。

在当前形势下,除了需要预防外部攻击,网络监测和管理部门也需要主动找出网络内部的异常,并且要能够快速发现引起内部异常的问题,并采取相应的安全措施。近年来网络流量异常检测与分析的常见技术有基于统计分析的检测技术[1]、基于神经网络的检测技术[2]、基于机器学习方法的检测技术[3]和基于代理的检测技术[4]。这些方法在预防网络攻击时各有优势,但也存在明显的缺陷[5]。机器学习方法通过学习固定的分类模式来检测外部的攻击,但该方法并不适合流量分析和网络内部异常发现,而且误报率高;神经网络方法需要利用大量的错误样例进行试错才能构建网络的初始拓扑结构和连接权值;代理检测技术通过数据的特征查找代理是否存在,该方法应用性较广,但在检测网络内部的异常时其检测效果一般。本文在文献[6-9]的基础上,提出了利用基于流的统计分析模型,对电力系统网络流的行为特征进行分析,从而达到实时监测外部攻击和内部异常的目的。本文基于大数据流的网络流量检测和分析系统,针对江苏省电力公司的营销、运监、生产与调度等业务场景,通过网络流量进行应用分析,取得了很好的实际效果。

1　系统架构和处理流程

1.1　技术架构

本文系统包括网络数据采集器、分布式实时数据传输通道、分布式流处理平台、网络数据协议特征库和大数据平台,系统技术架构如图 1 所示。

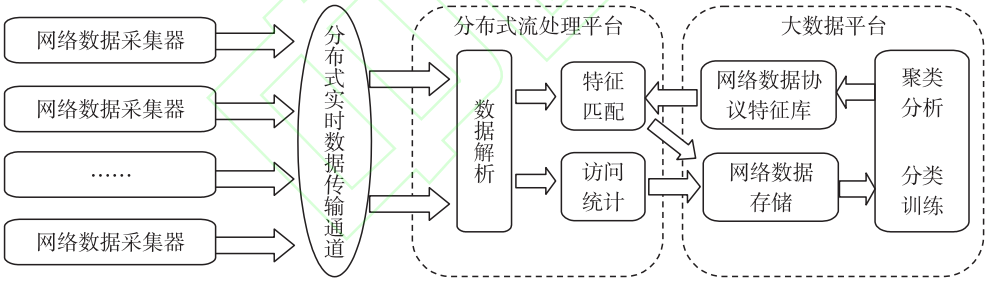

图 1　系统技术架构图

本文系统功能如下:

(1)利用分布式实时数据传输通道传送数据。网络数据采集器采集网络设备上的网络数据包,经分布式实时数据传输通道,实时发送队列化数据给分布式流处理平台。

(2)使用分布式流处理平台进行实时处理。分布式流处理平台对接收到的网络数据包进行实时数据解析,并通过网络数据协议特征库进行数据特征匹配,将经过匹配确认为异常的网络流量数据发送给大数据平台进行存储。

(3)大数据平台对存储的网络流量数据进行聚类分析、分类训练,并动态更新所述网络数据协议特征库。

1.2　系统处理流程

针对网络数据包分析结果,通过分布式实时数据传输通道,系统将采集到的数据实时发送到分布式流处理平台;分布式流处理平台对接收到的网络数据包进行实时数据解析,并通过网络数据协议特征库进行数据特征匹配,将经过匹配确认为异常的网络流量数据发送给大数据平台进行存储;大数据平台对存储的网络流量数据进行聚类分析、分类训练,并动态更新网络数据协议特征库[10]。处理流程可分为以下 4 部分:

(1)网络数据采集。在不影响程序正常运行的前提下,通过交换机把 1 个或多个端口的数据转发到某一个端口实现对网络的监听。网络数据

采集器的数量可动态扩展,可以通过增加网络数据采集器的数量增加网络采集数据的吞吐量。

(2)采集数据的实时传输。将采集到的数据实时地传送到分布式流处理平台,并利用分布式实时数据传输通道的队列化及可缓存的特性,实现采集数据的有序化、可扩容性管理。

(3)利用分布式流处理平台对数据进行实时处理。分布式流处理平台对网络数据包进行数据解析、特征匹配及访问统计。先将网络数据包根据协议信息进行实时解析,得到数据包的协议、发送地址、目的地址、发送端口、目的端口、数据包长度和数据包头部校验和;再根据解析的网络数据内容分别进行特征匹配和访问统计,特征匹配是将网络数据与网络数据异常特征信息通过匹配引擎进行异常判断,访问统计是通过统计一定时间内特定网络地址的所有访问进行异常判断;如果为异常网络数据,则将数据直接存储到大数据平台[11,12]。具体流程如图2所示。

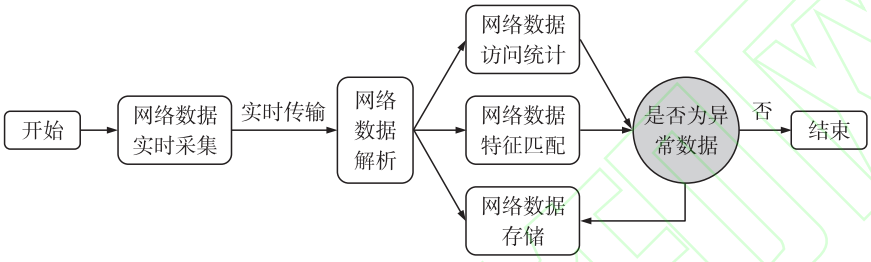

图2 检测方法流程图

在特征匹配时,采用恒虚警率异常探测算法(Reed-Xiaoli,RX)[13],该算法通过广义似然比检验检测异常值。设已知的正常网络数据特征为$\boldsymbol{x}_i$,n为已知的正常网络数据个数。首先计算已知的正常网络数据的均值和协方差矩阵的最大似然估计值$\boldsymbol{\mu}_z$、$\boldsymbol{C}_z$

$$\boldsymbol{\mu}_z=\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i \tag{1}$$

$$\boldsymbol{C}_z=\frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{x}_i-\boldsymbol{\mu}_z)(\boldsymbol{x}_i-\boldsymbol{\mu}_z)^{\mathrm{T}} \tag{2}$$

对未知的网络数据$\boldsymbol{y}$,计算其是否异常的RX算子的判别公式为

$$RX(\boldsymbol{y})=(\boldsymbol{x}-\boldsymbol{\mu}_z)^{\mathrm{T}}\boldsymbol{C}_z^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_z) \tag{3}$$

通过阈值t判断其是否为异常网络数据

$$\begin{cases}RX(\boldsymbol{y})\geqslant t & \text{正常}\\ RX(\boldsymbol{y})<t & \text{异常}\end{cases} \tag{4}$$

(4)网络数据协议特征库的动态更新。网络数据协议特征库包括网络流量数据基本特征和网络流量数据异常特征。根据协议类型,网络流量数据基本特征具体包括发送地址、目的地址、发送端口、目的端口、数据包长度和数据包头部校验,如表1所示。网络流量数据异常特征具体包括异常匹配表达式及相应的优先级。协议类型为IP、TCP[3]、UDP、ICMP协议。网络流量数据异常特征的异常匹配表达式是判断表达式,基于网络流量数据基本特征采用逻辑运算表达式描述异常行为,如表2所示。

表1 网络数据协议基本特征选项表

ID	特征选项	说明
1	协议类型	数据包采用的协议(TCP、UDP、ICMP)
2	发送地址	数据包源IP地址
3	目的地址	数据包目的IP地址
4	发生端口	数据包源地址的端口号
5	目的端口	数据包目的地址的端口号
6	数据包长度	数据包长度过大过小都可能是恶意的
7	数据包头部校验	对网络数据包的包头信息进行异常检测

表2 网络流量数据异常特征表

ID	协议类型	异常说明	异常判断表达式	优先级
1	IP	异常的TCP端口访问	*Value*(4)==30&&*Value*(5)==49	1
2	TCP	数据包流速过大	*Sum*(6,60)/60>70*1024*1024	2

表2中,当*ID*=1时,*Vaule*(4)表示校验网络流量数据的发生端口号,*Vaule*(5)表示校验网络流量数据的目标端口号。*Value*(4)==30&&*Value*(5)==49表示异常的TCP端口访问。当

$ID=2$ 时，$Sum(6,60)$ 表示对网络流量数据基本特征描述中 $ID=6$ 属性所对应的数值进行最近 60 s 内数据的求和，$Sum(6,60)/60$ 表示算出数据包最近 1 min 内的平均流速，异常判断表达式表示当平均流速>70 MB/s 时数据包流速过大。

(5) 基于模糊多标签支持向量基(Support vector machine，SVM)的网络数据分类。在网络数据存储结束后，本文利用模糊多标签 SVM 方法[14]对数据进行训练和分类。设 K 是类别个数，因此需要训练 K 个独立的 SVM 分类器，以适用于多类别数据的分类。设 p_i 表示第 i 个数据的特征，t_i 表示第 i 个数据的类别，则每个网络数据可用式(5)表示

$$\{(p_i,t_i,a_{i,k})\mid p_i\in I^d\}\quad a_{i,k}\in(0,1)\quad k=1,2,\cdots,K\quad i=1,2,\cdots,n \tag{5}$$

式中：$a_{i,k}$是第 i 个数据隶属于第 k 个类别的概率值，其求解形式如下

$$a_{i,k}=\begin{cases}1-|m_{+k}-p_n|/(r_{+k}+v) & t_i=+k\\ 1-|m_{-k}-p_n|/(r_{-k}+v) & t_i=-k\end{cases} \tag{6}$$

式中：$m_{+k}=\min\limits_i(p_i\mid t_i=+k)$，$m_{-k}=\min\limits_i(p_i\mid t_i=-k)$，$r_{+k}=\max\limits_{\{p_i:t_i=+k\}}|m_{+k}-p_i|$，$r_{-k}=\max\limits_{\{p_i:t_i=-k\}}|m_{-k}-p_i|$。

所有数据隶属于第 k 个类别的概率可表示成如下向量形式

$$\boldsymbol{a}_k=[\boldsymbol{a}_{1,k}\quad \boldsymbol{a}_{2,k}\quad \cdots\quad \boldsymbol{a}_{i,k}\quad \cdots\quad \boldsymbol{a}_{n,k}] \tag{7}$$

建立模糊多标签 SVM 的目标函数

$$\min_{\boldsymbol{w},\rho,\boldsymbol{b}}\frac{1}{2}\|\boldsymbol{w}\|^2+\boldsymbol{L}\boldsymbol{a}_i^{\mathrm{T}}\rho$$

$$\text{s.t.}\begin{cases}t_i(\boldsymbol{w}^{\mathrm{T}}p_i-b)\geqslant 1-\rho_i\\ \rho_i\geqslant 0\end{cases}\quad i=1,\cdots,n \tag{8}$$

用拉格朗日乘子法将上述带约束的优化问题转化为无约束优化问题，再通过其对偶形式进行求解，可得最终的模糊多标签 SVM 分类器。利用该分类器对测试网络数据进行分类，得到所有网络数据的类别。

2　系统实验与结果

为验证本文方法的有效性，设计了分布式安全联动网络，主要包括集群、大数据平台等设备，具体拓扑如图 3 所示。

在实际网络拓扑中，所有访问均通过顶层 2 台路由器接入，然后连接 2 台顶层交换机，对访问进行分发，至各个机房。系统实验需要对 2 台顶层交换机镜像端口数据进行采集，并对采集的网络数据进行解码、合并、排序，最后按照流或者按照业务逻辑过程，对数据包进行业务合成处理，计算流量、时延等信息，最终将采集信息及处理结果保存到大数据平台。

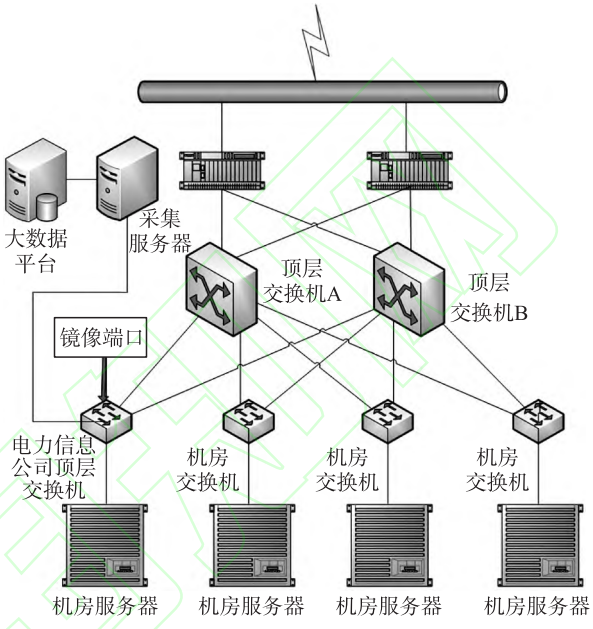

图 3　场景架构图

通过采集卡的处理，可将接收到的原始网络数据包转换为结构化的描述性信息，可直接供软件分析，输出的数据量大约为原始数据的 1%，信息量却得到完整的保留，降低了分析软件的负荷。采用嵌入式技术开发采集卡，将报文的接收、处理、计算、识别、输出合并在 1 块板卡上。所有处理都通过硬件实现，接入性能和处理能力都远远超过传统软件架构。采集卡主要功能架构图如图 4所示。

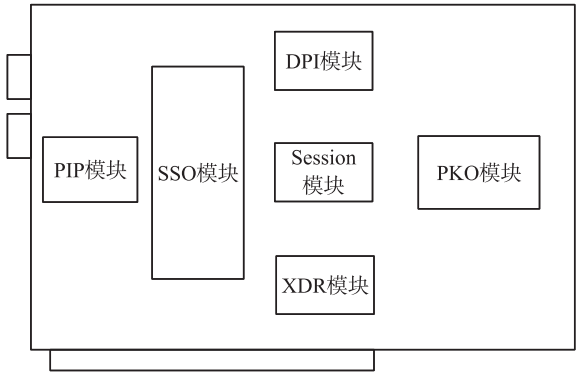

图 4　采集卡功能架构图

采集卡主要功能模块有 6 个，分别简述如下：(1) 原始信息处理(Primeval information

processor,PIP)模块负责接收报文,可支持多接口并发。

(2)单点登录(Single sign on,SSO)模块负责调度从PIP模块进入的报文,实现报文保序、优先级处理等功能。

(3)会话控制(Session)模块根据进入报文的五元组信息进行分组处理。

(4)高级处理接口(Deep process interface,DPI)对报文进行辨识,分析网络层到应用层的内容,识别应用类型。

(5)数据表示(X data representation,XDR)模块将前述各模块分析得到的信息组合成结构化的消息记录。

(6)公用输出(Public kinds data output,PKO)模块负责将结构化消息记录通过外部设备互联(Peripheral component interconnect,PCI)标准总线输出。

2.1 系统实施

通过对数据的抽取设计,开发了1套大数据平台数据抽取系统。该系统以Cloudera CDH为大数据平台的基础支撑软件,采用分布式流处理Spark Streaming平台对网络数据包进行数据解析、特征匹配及访问统计。该系统的集群由5个节点构成,每个节点配置24核CPU、127GB内存和10块磁盘。开发环境基于Eclipse,程序代码为Java,上层数据的查询与展现采用了Restful API,可以在浏览器中方便地发出查询数据的指令、实时地获取查询结果并展现在浏览器中。

在系统实现过程中,采用TcpDump采集数据,通过配置交换机把1个或多个端口的数据转发到某一个端口,进而实现网络监听,不影响正常程序的运行;同时采用Apache Kafka对采集的数据进行实时传输,将数据发送到Spark Streaming分布式流平台实时处理,将分析的结果存储在大数据平台,通过大数据平台的MapReduce、Hive、Mahout等组件使用支持向量机和贝叶斯构造的分类器进行分类训练,对网络数据协议特征库进行更新。

(1)环境搭建。系统验证环境搭建如表3所示。

表3 系统验证环境配置表

分类	标号	设备	用途
硬件部分	1	采集卡1块	用于网络流数据采集
硬件部分	2	X86服务器1台	用于安装采集卡
硬件部分	3	千兆转万兆转换设备1台	测试环境为千兆接口,为生产环境准备的采集卡为万兆接口,临时增加转换设备
硬件部分	4	光纤尾纤若干,千兆光模块2个,万兆光模块2个	用于设备间的连接
软件部分	5	采集卡驱动程序1套	

(2)场景架构。系统对网络核心交换机的端口镜像数据进行采集监测。网络核心交换机的端口镜像数据采用1 000 M光纤输出,再通过千兆转万兆设备将信号转换为10 G光纤信号,输入采集卡,经过采集卡处理后,输出结构化网络流数据,存储在X86服务器上,定期通过约定接口将数据推送至大数据平台上。系统场景架构图如图5所示。

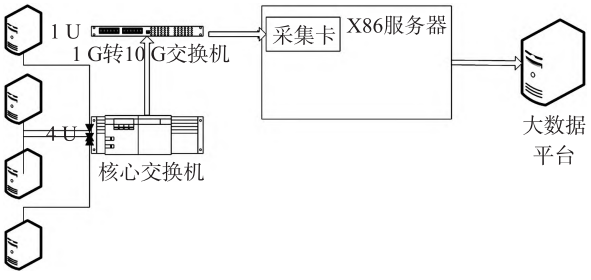

图5 场景架构图

(3)实施步骤。系统测试环境的部署及准备工作具体如下:将采集卡安装到X86服务器上;将X86服务器部署到机房与核心交换机相邻的机柜,便于布线工程实施;将核心交换机镜像口与千兆转万兆设备的千兆入接口之间用光纤连接千兆光模块;在千兆转万兆设备的出接口和采集卡的入接口之间使用光纤连接万兆光模块;配置X86管理网口,在X86服务器上安装采集卡驱动程序;请求机房管理员启动核心交换机的镜像功能。

2.2 参数选择和结果对比

在计算网络流数据的特征匹配时,RX算法中的阈值选择对异常数据的检测正确率有较大的影响。为此,在区间[1,10]中选择10个阈值,从训练数据集中选取30个正常网络数据作为已知的正常网络数据集,选取10个正常网络数据和

10 个异常网络数据进行测试,计算不同阈值下其异常准确率,结果如图 6 所示。从图 6 中可以看出,随着阈值增大,正常网络数据的检测正确率逐渐降低,异常网络数据的检测正确率不断提升,而整体网络数据的检测正确率在阈值为 8 时达到峰值,这表明选取阈值 $t=8$,异常检测的结果最为准确,因此在后续实验中,令 $t=8$。

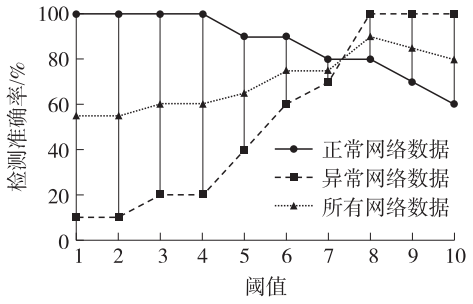

图 6　阈值对网络数据检测准确率的影响曲线图

为比较分类的准确性,选 1 100 组数据构成数据集,其中随机选择 100 组作为训练样本,其余数据作为测试样本。2 种非监督聚类方法为模糊 C 均值聚类(Fuzzy C means,FCM)、高斯混合模型(Gaussian mixture model,GMM),3 种监督分类方法为 K 近邻(K nearest neighbor,KNN)、SVM 以及模糊 SVM 分类方法。分别采用这 5 种方法对测试样本进行分类,并计算正确性指标

$$J=\frac{N_c}{N_t}\times 100\% \tag{9}$$

式中:N_c 是被正确分类的样本个数,N_t 是测试样本总个数。不同分类方法所得的分类准确性如图 7 所示。从图 7 中可以看出,与非监督聚类方法相比,监督分类方法所得分类的准确性较高;与其他方法相比,模糊 SVM 方法所取得的正确性最高。

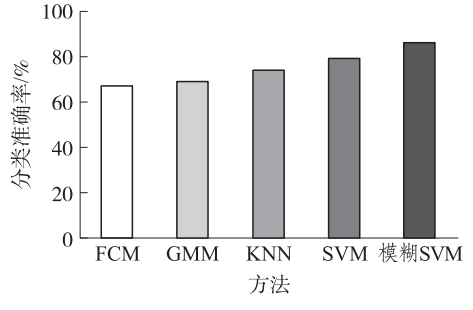

图 7　不同方法的分类准确率比较图

2.3　实验结果分析

由图 7 结果可知,在采用基于大数据的数据采集方法进行数据采集时,正确率高,误报率低;同时,本文方法的分类准确率和耗费时间等指标都得到明显提高。但当网络流量发生异常时,流量属性的统计特性变化十分微小,很难通过这些微小变动检测异常,需要对异常检测算法进行优化提高。

(1)在网络流数据处理中,从交换机镜像口采集数据,计算网络层流量,当 TCP 发生重传时,同样会统计到网络层流量中。服务器软件统计的是应用层流量,不包含重传流量。当网络状况变差发生 TCP 重传时,两者统计的数值会有少量误差。

(2)采集机和服务器的系统时间会有不同步的情况,例如采集机时间为 11:59,服务器时间为 12:00,此时对服务器 12:00 收到的报文采集机会标示为 11:59 收到,在这种情况下,按照服务器时间统计 12:00 ~ 13:00 的数据就会产生误差。

(3)对较长的时间不同步,可通过修正查询范围解决。如服务器指定 12:00 ~ 13:00,而采集数据指定 11:59 ~ 12:59,可降低误差。但由于人工观察有误差,不同步的时间很难精确到 s,因此对比两者的统计结果会有少量的误差。

3　结束语

针对网络流量异常检测的问题,传统的解决方法有基于规则、统计分析、有限状态机等方法,这些方法能够在一定的程度上检测流量异常,但在处理数据量、实时性等方面存在一些不足。针对以上问题,提出了 1 种基于大数据流的网络流量检测与分析方法,并开发了 1 套基于 Cloudera CDH 的系统,实现了大数据平台海量用电数据的高效抽取。该系统采用分布式流式处理机制达到实时检测,并且利用大数据平台分布式存储、数据计算分析的能力,实现网络数据分布式存储,准确训练网络数据协议特征库。该系统较好地利用了大数据平台优势,提高了网络流量异常检测的效率,具备处理并检测海量异常数据的能力,在网络流量异常特征检测时具备可扩展性、可配置性,在计算能力上具备线性扩展性能力。在江苏省电力公司的营销、运监、生产与调度等业务场景下,本系统通过网络流量分析,取得了很好的实际效果。

参考文献:

[1] Ho C Y, Lai Y C, Chen I W, et al. Statistical analysis of false positives and false negatives from real traffic with intrusion detection/prevention systems[J]. IEEE Communications Magazine, 2012, 50(3): 146-154.

[2] Chen Xueyun, Xiang Shiming, Liu Chenglin, et al. Vehicle detection in satellite images by hybrid deep convolutional neural networks[J]. IEEE Geoscience and Remote Sensing Letters, 2014, 11(10): 1797-1801.

[3] Mukesh K G, Khanna H P, Velvizhi R V. An anamoly based intrusion detection system for mobile ad-hoc networks using genetic algorithm based support vector machine [J]. Advances in Natural and Applied Sciences, 2015, 9(12): 40-45.

[4] Khaled O, Marín A, Almenares F, et al. Analysis of secure TCP/IP profile in 61850 based substation automation system for smart grids [J]. International Journal of Distributed Sensor Networks, 2016, 2: 1-11.

[5] 王元卓,靳小龙,程学旗. 网络大数据:现状与展望[J]. 计算机学报,2013,36(6):1125-1138.
Wang Yuanzhuo, Jin Xiaolong, Cheng Xueqi. Network big data: Present and future [J]. Chinese Journal of Computers, 2013, 36(6): 1125-1138.

[6] Thimma M, Liu F, Lin J Q, et al. HyXAC: Hybrid XML access control integrating view-based and query-rewriting approaches [J]. IEEE Transactions on Knowledge and Data Engineering, 2015, 27(8): 2190-2202.

[7] 臧天宁,云晓春,张永铮. 网络设备协同联动模型[J]. 计算机学报,2011,34(2):216-228.
Zang Tianning, Yun Xiaochun, Zhang Yongzheng. A model of network device coordinative run[J]. Chinese Journal of Computers, 2011, 34(2): 216-228.

[8] Li Yuchong, Luo Xingguo, Li Bainan. Detecting network-wide traffic anomalies based on robust multivariate probabilistic calibration model [C]// Military Communications Conference, MILCOM 2015. New York, NY, USA: IEEE, 2015: 1323-1328.

[9] 黄伟,陈昊,郭雅娟. 融合领域知识的网络异常检测方法[J]. 南京理工大学学报, 2016, 40(2): 229-235.
Huang Wei, Chen Hao, Guo Yajuan. Network anomaly detection approach using domain knowledge [J]. Journal of Nanjing University of Science and Technology, 2016, 40(2): 229-235.

[10] 郑琨琪,何光宇. 智能用电网络数据采集与通信机制的研究[J]. 中国电机工程学报,2016,36(6): 1544-1551.
Jia Kunqi, He Guangyu. Research of smart electric appliance network data collection and communication mechanism[J]. Proceeding of the Chinese Society of Electrical Engineering, 2016, 36(6): 1544-1551.

[11] Gao Yun, Fu Xiao, Luo Bin, et al. Haddle: A framework for investigating data leakage attacks in hadoop[C]// 2015 IEEE Global Communications Conference (GLOBECOM). New York, NY, USA: IEEE, 2015: 1-6.

[12] Box G E P, Jenkins G M, Reinsel G C, et al. Time series analysis: Forecasting and control [M]. 5th ed. New Jersey, USA: Wiley, 2015.

[13] Zhou J, Kwan C, Ayhan B, et al. A novel cluster kernel RX algorithm for anomaly and change detection using hyperspectral images [J]. IEEE Transactions on Geoscience and Remote Sensing, 2016, 54(11): 6497-6504.

[14] Wang Shuihua, Yang Xiaojun, Zhang Yudong, et al. Identification of green, oolong and black teas in China via wavelet packet entropy and fuzzy support vector machine[J]. Entropy, 2015, 17(10): 6663-6682.