[理工科技术应用研究]

DOI 10.16221/j.cnki.issn1671-1084.2017.03.026

基于数据流聚类算法的网络异常检测系统设计

莫徽忠

（柳州职业技术学院，广西　柳州　545006）

摘要：为保障校园网络信息安全和提高其带宽使用效率，采用 sniffer 流量方法采集数据，通过 CluStream 聚类算法进行聚类处理，实现对校园网络数据流的异常检测。通过对端口数据流进行检测，表明系统能够发现异常数据并给出预警信息。

关键词：聚类算法；网络数据流；网络安全

中图分类号：TP393.08　　文献标志码：A　　文章编号：1671-1084（2017）03-0099-05

0 引言

近几年，各高校为了满足教育信息化要求，都加强了校园网络建设，校园网的环境得到改善。但随着教育技术的发展以及云桌面的使用，教学对网络的依赖也越来越强，如何保障好校园网络的信息安全和提高其使用效率，成为亟需解决的问题。这需要进行网络中异常数据的监测，及时发现端口扫描、dos 攻击、大规模蠕虫病毒等对 TCP/IP 协议和网络物理链路带来不安全的威胁，以及个别用户滥用网络资源[1]的情况。由此可见，网络异常检测系统的设计是一个值得探讨的问题。

1 系统各模块及其主要功能

系统主要由四大模块组成。第一模块完成数据采集及预处理功能，主要采用 sniffer 流量方法采集数据和混合指数直方图进行预处理，获取概要信息并存入信息库中；第二模块可实现聚类挖掘功能，对信息库中的数据进行聚类，得到频繁模式并存入数据库中；第三模块具有异常检测分析功能，对频繁模式进行比较分析，从而发现异常数据流，将异常检测情况反馈给异常告警模块；第四模块完成异常告警处理，根据异常情况对频繁模式数据库进行更新或进行应急处理。

2 系统实现

2.1 数据采集及预处理

在众多的数据中，校园网络中的数据能够体现用户行为的数据才是需要重点采集数据。把用户的 IP 地址与 MAC 地址捆绑是校园网络常用的方法，因此要找到异常用户，只要找到 MAC 地址或 IP 地址，再由其对应关系即可找到。检测网络数据是否异常，可以通过检测源 IP 的网络流量，包括其所接到同一目的 IP 的流量，还包括检测相同发送数据的 IP 与相同接收数据的 IP 连接频率是否正常。校园网网络异常网络的特征值可以选择数据发起 IP 与数据接收 IP 的地址、网络数据连接流量、数据流发起时间和数据流终止时间、同一发送数据的 IP、同一接收数据的 IP 连接频率，而源 IP 与宿 IP 地址连接频率可以归类到构造属性中，其他特征值归类到原始属性中[2]。

收稿日期：2017-05-04

基金项目：2014 广西高校科学技术研究（LX201453）

作者简介：莫徽忠，硕士，柳州职业技术学院副教授，通信与网络技术团队负责人，研究方向为通信与网络技术。

数据采集可以基于时间,也可以基于数据包,而网络数据采集常用的有四种,有基于硬件探针、sniffer流量、SNMP流量以及Netflow流量[3]。基于sniffer流量方法可以获取较全面的数据流信息,因此采用这种采样技术。

由于校园网络数据流具有突发性,有很多数据并不是一致的也不完整,这些数据并不能立即用于后续处理,应该先进行预处理,一方面将确定性数据通过相似性转化变成了非确定性的数据,达到了保护隐私的目的;另一方面就是选择适当的网络行为特征值,将其合并和离散化,便于后面聚类处理。为了使数据完整可以采用K—means聚类算法填充数据流缺失值。

概要数据信息的获取有抽样、Hash方法、直方图和小波分析等方法。直方图有等宽和指数直方图,或者将两者进行混合,构成混合指数直方图。混合指数直方图再进一步划分为基于时间的桶,就可以把直方图与数据值的范围和时间范围联系在一起。用Wi表示数据流时间权值,数据流到达较早则Wi大,数据流到达较晚Wi就小;TAi和WAi分别表示活动直方图平均时间戳与平均权值。其过程分为两步:第一步,初始化的混合直方图,起始滑动窗口内的数据可以采用聚类处理,得到数据模型后放在不同的桶内,桶与不同的混合直方图保持对应关系,每一个初始直方图也随时变化;第二步,直方图的实时更新,直方图的变化是随新数据不断到达而不断改变,并判决该桶过期与否,统计活动直方图个数,并与滑动窗口长度的极大值进行比较,条件满足则选出要更新的活动直方图进行更新,否则,合并整理邻接的活动直方图,实现活动直方图更新。

2.2 聚类挖掘

对网络数据统计信息进行更新与维护就要增加时间特性,实现时间维度的扩展,并对进行CF加法定义。

定义1(聚类特征) 网络数据异常检测的聚类特征(CF，Clustering Feature)定义可以描述为包含d维数据集合(…，$\overrightarrow{X_{i-1}},\overrightarrow{X_i},\overrightarrow{X_{i+1}}$,…)的聚类信息的(2d+2)元组，即设给定一个子簇中的 d 维数据集合$\{\overrightarrow{X_i}|0<i\leq n\}$，则有聚类特征

$$\overrightarrow{CF}=(\vec{S},\vec{D},n,t)$$

其中，$\vec{S}=[\sum_{i=1}^{n}x_i^1,\sum_{i=1}^{n}x_i^2,\cdots,\sum_{i=1}^{n}x_i^d]$;

$\vec{D}=[\sum_{i=1}^{n}(x_i^1)^2,\sum_{i=1}^{n}(x_i^2)^2,\cdots,\sum_{i=1}^{n}(x_i^d)^2]$;

n表示子簇中数据的数量；t表示该特征失量的存储时刻。

当有新的网络数据输入，利用CF加法操作把现有子簇与其合并，实现网络数据统计信息的更新与维护。

定义2(CF加法)设网络数据异常检测模型中，初始时刻t_1子簇C_1的聚类特征矢量为$\overrightarrow{CF_1}=(\vec{S_1},\vec{D_1},n_1,t_1)$，$t_2$时刻子簇$C_2$的聚类特征矢量为$\overrightarrow{CF_2}=(\vec{S_2},\vec{D_2},n_2,t_2)$且$C_1\cap C_2=\varphi$，则两个子簇合并后，其合并子簇信息的CF特征矢量为

$$\overrightarrow{CF_\Sigma}=\overrightarrow{CF_1}+\overrightarrow{CF_2}=(\vec{S_1}+\vec{S_2},\vec{D_1}+\vec{D_2},n_1+n_2,t)$$ t取t_1和t_2中的最大者。

CF加法操作在新数据输入后,完成两个子簇合并,即对网络数据统计信息得到及时准确的存储和维护,使网络数据统计信息满足后面聚类分析的要求。

随新数据流不断输入,通过加法操作进行聚类,对网络数据统计信息以足够细的粒度维护并更新,保证同一子簇中相同类型数据点的唯一性,满足异常检测对攻击检测的要求。算法开始,各子簇集合初始化后均为空集,随着数据的到达,在聚类挖掘后生成不同的子簇。当新的数据点到达后,算出该数据点到各子簇中心的最短距离,将它与设定子簇半径的阈值进行比较,根据比较结果选择合并到现有子簇,还是创建新的子簇;要创建新子簇必须是总子簇数小于极大值的条件下;若不大于,就将该数据点合并到距离最近的子簇。

2.3 异常检测

异常检测是在输入数据概要信息的 CF 特征矢量集合之后，将其与不同时间段的 CF 特征矢量集进行聚类分析，完成数据点归类。CF 特征矢量集合的 CF 减法操可对特定时间窗口的网络数据进行计算，使数据统计信息的 CF 特征矢量集合实时反映相应时间段内网络数据的实时变化。

定义 3(CF 减法) 设网络数据异常检测模型中，初始时刻 t_1 子簇 C_1 的聚类特征矢量为$\overrightarrow{CF_1} = (\vec{S}_1, \vec{D}_1, n_1, t_1)$，$t_2$时刻子簇 C_2 的聚类特征矢量为$\overrightarrow{CF_2} = (\vec{S}_2, \vec{D}_2, n_2, t_2)$

且 $t_1 < t_2$，$C_1 \subseteq C_2$，对于(t_1 t_2]时间窗口内的网络数据集，两个子簇的 CF 特征矢量有

$$\overrightarrow{CF}_{\Delta} = \overrightarrow{CF_2} - \overrightarrow{CF_1} = (\vec{S}_2 - \vec{S}_1, \vec{D}_2 - \vec{D}_1, n_2 - n_1, t\ )$$

其中 $t=t_2-t_1$。

通过对选择时间窗口中到达数据的聚类特征矢量进行减法操作来实现异常检测，在不考虑以前的历史数据的情况下，实现对某一特定时间窗口的网络数据情况进行分析，从而准确体现了网络的数据流现状，完成对网络异常行为检测。即通过减法操作计算近期网络数据信息，实现对当前网络的攻击行为的检测。

由于输入的数据是 CF 特征矢量，它是多个数据点集合的描述。为了充分利用了现有的 CF 特征矢量信息，在异常检测算法中，利用全部数据点信息的簇间距离计算方法，计算子簇内全部的数据点：

设输入子簇 $SubC_k$，$CenC_l$ 为代表簇 C_l 中心的子簇，N_1 是 $SubC_k$ 中的数据点数目， N_2 为 $CenC_l$ 中的数据点数目，$\vec{X}_i$ 表示 $SubC_k$ 中的 d 维数据点，$\vec{Y}_j$ 表示 $CenC_l$ 中的 d 维数据点，$i \in [1, N_1]$，$j \in [1, N_2]$，那么子簇 $SubC_k$ 到簇中心 $CenC_l$ 的距离可以下面的式子求得

$$dist(SubC_k, CenC_l) = \sqrt{\frac{1}{N_1 N_2} \sum_{i=1}^{N_1} \sum_{j=1}^{N_2} (\vec{X}_i - \vec{Y}_j)^2}$$

将上式中等号右边展开，即有

$$dist(SubC_k, CenC_l) = \sqrt{\frac{1}{N_1 N_2} [N_2 (\sum_{i=1}^{N_1} \vec{X}_i^{\,2}) + N_1 (\sum_{j=1}^{N_2} \vec{Y}_j^{\,2}) - 2 (\sum_{i=1}^{N_1} \vec{X}_i)(\sum_{j=1}^{N_2} \vec{Y}_j)}$$

上式中，$\sum \vec{X}_i^{\,2}$ 与 $\Sigma \vec{Y}_j^{\,2}$ 分别为描述两个子簇的 CF 特征矢量的$\vec{D}$，而 ΣX_i 与 ΣY_j 分别为两个子簇的 CF 特征矢量的$\vec{S}$，因此将 CF 特征矢量代入上式即可计算求得$dist(SubC_k, CenC_l)$。

2.4 告警处理

在得到异常检测结果之后，系统根据异常情况对频繁模式数据库进行更新或进行应急处理。

3 验证分析

数据来自某校园网络安全监测平台。采用滑动时间窗口技术[4]，采集 360 个端口及其流量值，对排名前 6 位的端口及其流量变化情况进行数据分析。如图 1 所示，在所统计的 24 小时内，6 个端口的流量数据均表现正常；如图 2 所示，是某网站蠕虫爆发时的网络数据，72 端口流量值随时间变化情况，表明流量数据明显异常。

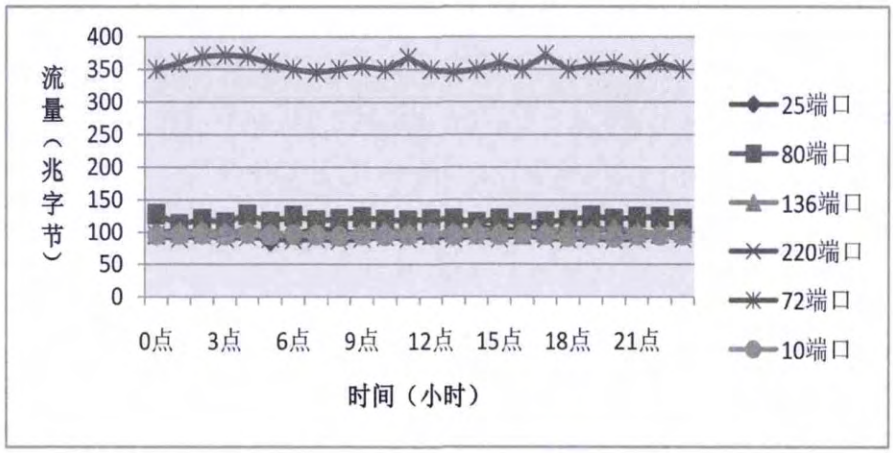

图 1　24 小时流量前六名端口流量情况

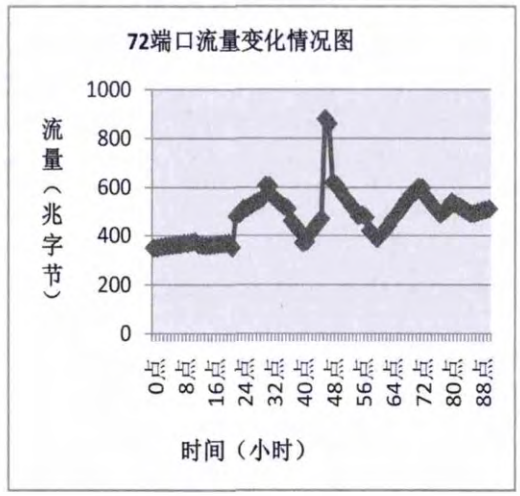

图 2　72 端口流量变化情况

通过聚类挖掘实现对网络数据流的聚类，得到正常和异常情况下数据流频繁模式，表 1 和表 2 分别是部分正常频繁模式和部分异常端口频繁模式。

表 1　部分正常的端口频繁模式

时间	次数	支持数	（端口，离散值）
14261100	50	6	（25，3）（80，40）（136，3）（120，2）（35，2）（260，2）
14261200	45	6	（25，4）（80，30）（136，3）（120，2）（35，2）（260，2）
14271200	47	6	（25，3）（80，50）（136，3）（120，3）（35，2）（260，2）
14271300	40	5	（260，2）（130，2）（246，3）（310，2）（48，2）
14271100	40	5	（130，3）（246，2）（310，3）（48，3）（56，3）

表 2　部分异常的端口频繁模式

时间	次数	支持数	（端口，离散值）
14281900	1	6	（25，3）（80，40）（136，3）（72，5）（35，2）（260，2）
14282200	1	6	（25，4）（80，45）（136，3）（72，6）（35，2）（260，2）
14292400	1	6	（25，4）（80，50）（136，4）（72，7）（35，2）（260，2）
14290500	1	6	（25，4）（80，50）（136，3）（72，7）（35，2）（260，2）
14290600	1	6	（25，4）（80，45）（136，3）（72，4）（35，3）（260，2）

网络数据流异常检测时，应根据实际情况设定权值和阈值。这些参数的选取关系到检测异常的及时性，通常以实验来获取较好权值和阈值的取值，选择不同权值和阈值实现系统检测的准确性。检测结果如表 3 所示。

表 3　异常检测结果

权值 \ 时间		14281900	14282200	14292400	14290500	14290600
Wi=1	检测值（阈值 d=10.5）	2.3	10.2	10.7	10	1
	检测情况	No	No	Yes	No	No
Wi=0.5	检测值（阈值 d=2.5）	1.6	3	4.6	2.0	1
	检测情况	No	Yes	Yes	No	No

4 小　结

本数据流异常检测系统采用 sniffer 流量方法采集数据和混合指数直方图进行预处理，通过 CluStream 聚类算法进行聚类，对数据流进行异常检测和分析，实现基于聚类算法的数据流异常检测系统。通过实验表明，本系统能够对数据流异常检测，可以实现对异常使用用户的监控，满足校园网络安全监测要求。

参考文献：

[1]刘秋兰.基于流数据挖掘的网络行为分析及应用研究[D].苏州：苏州大学，2008.

[2]李凌，李玲玲. 校园网络异常行为检测技术研究与应用–基于数据流的网络数据采集技术[J]. 新余学院学报，2015,20(3):15–18.

[3]黄超.网络异常行为检测与分析方法研究[D].西安：西安电子科技大学，2010.

[4]Zhu Y,Shasha D. StatStream: Statistical monitering of thousands of data streams in real time [C]//Proc. of the28th Int'l Conf. on Very Large Data Bases. Hong Kong: Morgan Kaufmann, 2002:358–369.

System Design on the Network Anomaly Detection Based on Data Stream Clustering Algorithm

MO Hui–zhong

(Liuzhou Vocational & Technical College, Liuzhou Guangxi 545006, China)

Abstract: To protect the campus network information security and improve the bandwidth efficiency, sniffer flow method is used to collect data, and CluStream clustering algorithm is used to cluster the data of campus network. By detecting the port data stream, it indicates that the system can detect the abnormal data and give the warning information.

Key words: clustering algorithm; network data flow; network security