第46卷 第4期 计 算 机 工 程 2020年4月
Vol.46 No.4 Computer Engineering April 2020

·网络空间安全· 文章编号: 1000-3428(2020)04-0135-08 文献标志码: A 中图分类号: TP309

基于用户窗口行为的内部威胁检测研究

李 志,宋礼鹏

(中北大学 大数据学院 大数据与网络安全研究所,太原 030051)

摘 要: 用户在计算机上的行为直接体现在与应用窗口的交互过程中。针对内网安全问题,从应用窗口的使用角度出发,对用户行为进行研究。搭建完全自由的内网环境,采集与分析用户在应用窗口上的行为数据,提取面向异常用户检测与用户变化行为识别的行为特征。通过样本均值分布特性和K-S检验验证了不同用户使用应用窗口的行为存在显著差异,并结合欧氏距离与置信区间,构建异常行为检测算法。实验结果表明,该算法能够有效检测异常用户与识别用户变化行为,准确率分别高达97.4%和94.5%,对于内部威胁防御具有重要作用。

关键词: 内网安全; 应用窗口; 用户行为; 异常检测; 欧氏距离

开放科学(资源服务)标志码(OSID):

中文引用格式: 李志,宋礼鹏. 基于用户窗口行为的内部威胁检测研究[J]. 计算机工程,2020,46(4):135-142,150.

英文引用格式: LI Zhi, SONG Lipeng. Research on internal threat detection based on user window behavior[J]. Computer Engineering,2020,46(4):135-142,150.

Research on Internal Threat Detection Based on User Window Behavior

LI Zhi,SONG Lipeng

(Research Institute of Big Data and Network Security,School of Big Data,North University of China,Taiyuan 030051,China)

【Abstract】 User behavior on a computer is directly reflected in the interactions with application windows. To address intranet security issues,research on user behavior is conducted from the perspective of the use of application windows. A completely free intranet environment is built,and user behavior data on application windows is collected and analyzed. On this basis,two kinds of behavior features of the use of application windows are extracted,which solve abnormal user detection and user change behavior recognition respectively. By using the sample mean distribution features and K-S test, it is verified that there are significant differences in the behavior of different users using application windows. Then,an abnormal behavior detection algorithm is constructed by combining Euclidean distance and confidence interval. Experimental results show that the algorithm can detect abnormal users and identify changed user behavior with a high accuracy. The accuracy rates are 97.4% and 94.5% respectively,which has practical application significance for preventing internal threats.

【Key words】 intranet security; application window; user behavior; abnormal detection; Euclidean distance

DOI: 10.19678/j.issn.1000-3428.0055801

0 概述

随着科学技术的发展,企业信息化水平不断提高,对企业和组织的信息安全管理和分析提出了新的挑战,其中来自内部的威胁不断增加,并给企业和组织造成巨大损失。对于内部威胁而言,内部恶意用户具有合法的资源访问权限,因此可通过合法的操作发起攻击并且很难被现有的访问控制或身份认证机制进行检测。一旦内部攻击发生,就会对内部资产造成巨大损失[1-3]。因此,如何有效防御内部威胁一直是研究的热点[4-5]。

文献[6]根据Unix系统环境中的Unix Shell命令段序列的使用特征来描述用户行为,如命令的名称、单条命令的执行时长、CPU运行时长和内存占用。该研究主要是针对Unix系统的用户行为,而相比Unix系统,Windows操作系统的使用更加广泛,并且在Windows环境下用户主要是基于窗口界面与计算机进行交互,很少使用命令行进行操作。因此,对Unix用户的分析技术不能完全适用于Windows系统环境。文献[7]分析Windows系统API与进程

基金项目: 国家自然科学基金(61772478)。

作者简介: 李 志(1992—),男,硕士研究生,主研方向为大数据分析、网络安全、数据挖掘; 宋礼鹏,教授、博士。

收稿日期: 2019-08-23 修回日期: 2019-10-08 E-mail: 2322257662@qq.com

表信息,提出监控系统调用方法,通过用户、文件与进程的关联关系建立文件访问与进程调用的联系,不足是仅能检测缓冲区溢出,然而依据缓存区独立研究其不能完全表征用户在计算机上多样性的操作行为。文献[8]通过记录用户在使用某特定程序时 Windows Native API 的调用序列,利用隐马尔科夫模型对正常用户行为进行建模,仅利用 3 台主机模拟内网环境,并且只将 Windows 系统中的 4 种进程作为异常发生时的测试对象。另外,内网中异常用户在计算机上执行的许多操作在技术上是合法的,并不会造成进程异常。文献[9]通过收集和分析 Windows NT 系统的审计日志,提出一种用户行为建模方法,使用 SVM 分类器识别异常行为。虽然审计日志记录了用户的操作数据,如应用程序窗口名、应用程序开始和结束时间及对应的进程 ID 等,但因有限的属性记录以及简单拼接不同类型的审计数据会造成特征失效、模型训练复杂度过高和模型过拟合问题,并且实验结果准确率和误报率不能满足实际要求,当准确率为最高的 66.7% 时,误报率达到 11%,当误报率为 3.7% 时,准确率为 63%。文献[10-11]提出基于文件使用的内部威胁检测系统,该系统用于伪装者攻击,从用户遍历文件系统以及访问文件目录的角度建立行为模型,其局限是仅针对文件系统单一域访问的异常情况进行研究。文献[12]收集了不同用户调用窗口的标题和进程等信息,但由于数据获取难度大,因此在此方面很难取得重大进展。

内部威胁的产生原因主要包括计算机在无人看管时,他人进行一些不合法的操作,以及计算机真实用户的行为发生恶意变化后的操作。无论在何种情形下,用户在计算机上的行为直接体现在与应用窗口的交互过程中。针对内部威胁的产生原因,本文从应用窗口的使用角度分别研究异常用户检测和用户自身变化行为的识别,提出可表征用户窗口行为的特征。借助样本均值分布特性和 K-S 检验,结合采集的用户行为数据以验证用户窗口行为的差异性,并利用欧氏距离和置信区间构建异常检测算法。

1 用户窗口行为特征提取

针对异常用户检测和用户自身变化行为识别问题,通过分析用户窗口的行为习惯分别提出两类窗口行为特征和两种行为模式建立方法。对于每类行为特征,根据用户行为数据分析特征值范围的统计分布来进行行为度量,可以有效地量化用户窗口使用行为的差异。

1.1 用户行为特征分类

用户在一个应用窗口内的活动有两种模式: 第一种是用户在窗口内间断的活动,即用户的动作之间存在间隔,这些间隔由用户脱机行为造成,一个窗口置顶的时间越长越能体现出这种行为模式,将连续动作的时间长度称为窗口内动作的有效时长,用户的脱机时间长度称为动作间隔时长,窗口内的有效动作指的是鼠标和键盘动作; 第二种是用户在一个窗口内的动作结束后随即离开当前窗口。

实验主要提取了置顶时长、单个活动时长均值和方差、有效总时长、动作间隔时长均值和方差、间隔总时长、间隔次数、鼠标和键盘动作时长、单个有效活动时长和间隔时长的均值与方差这些窗口内的行为特征。

图 1(a)和图 1(b)分别显示了用户在窗口内单次活动持续平均时长分布和活动总持续平均时长分布,均是重尾分布。其中单次活动的平均时长分布主要集中在 5 min 内。在 0~3 min 内,用户单次活动平均时长有明显的波动; 活动总持续平均时长超过 90% 的时长小于 10 min,在 0~5 min 内,较高的误差线表明不同用户通常具有不同的有效活动时长。

图 1(c)显示了用户活动间隔时长的差异。为清晰地呈现用户行为之间的差异,本文对 5 个用户进行刻画。从窗口内活动间隔数据经过 Box-Cox 变换[13]处理后得到的概率密度图可以直观看出,用户在窗口内的活动间隔时长分布存在显著差异。

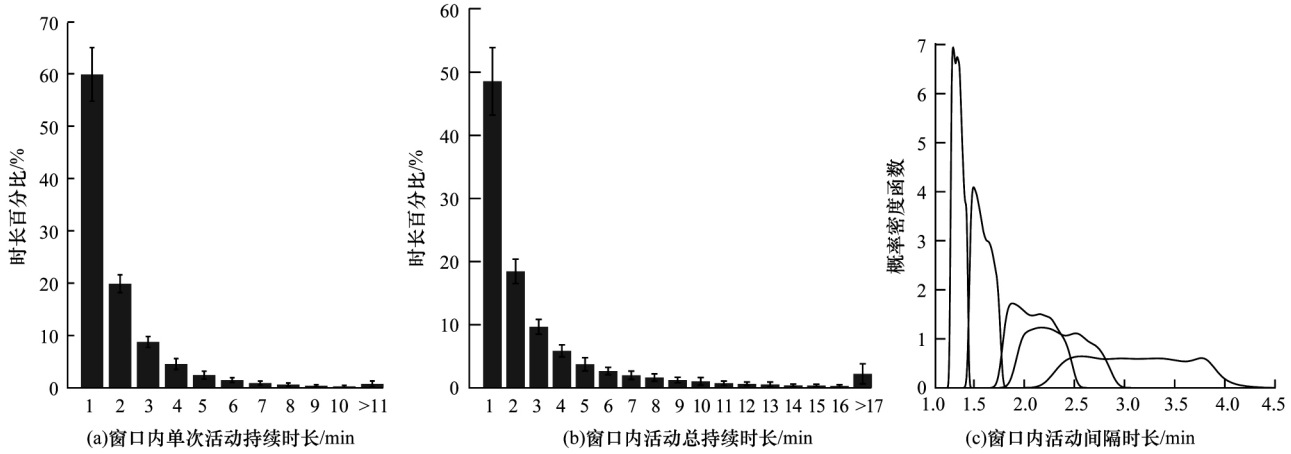

图 1 窗口内部行为特征分布

Fig. 1 Distribution of behavior characteristics inside the window

窗口大小、切换顺序等行为也是表征用户使用应用窗口的一类关键特征。下文详细描述这些特征:

1) 窗口名: 用户在一段时间内对不同功能软件的需求是稳定的,此特征可体现用户对软件的使用偏好。图 2 显示了所有用户各类窗口平均时长占比的分布情况。可以看出,用户在外部浏览器、Office、论文阅读软件、桌面、编程软件和即时通讯的平均耗时占所有窗口时长的 90% 左右。其中外部浏览器、Office 和论文阅读软件占相当高的比例。非常高的误差线表明用户对各类型的窗口活动呈现出多样性和差异性。

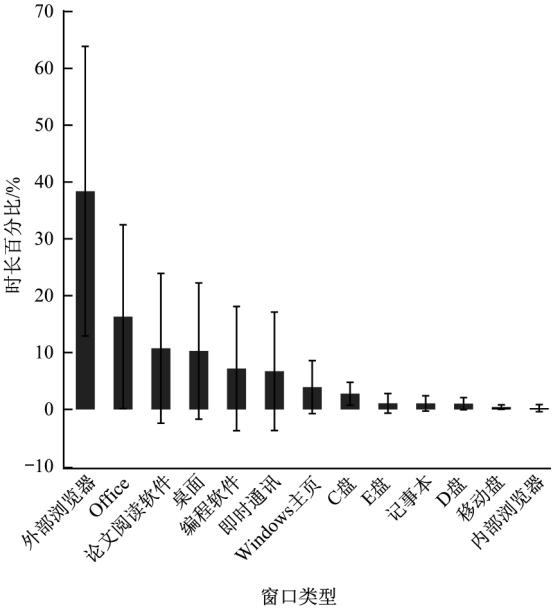

图 2　各类用户窗口总持续平均时长分布

Fig. 2　Distribution of total duration of each type of window

2) 活跃状态的窗口数: 活跃状态指的是所有打开的窗口,包含最大化、最小化和特定大小的窗口。用户在窗口内的动作结束后对当前窗口有 3 种处理方式,即关闭、最小化和覆盖当前窗口。此特征能够直接反映用户对窗口状态控制的习惯。

3) 最小化、最大化窗口数: 结合活跃状态窗口数,此特征能够更加细致地描绘用户对窗口的控制习惯。

4) 当前工作窗口状态: 在用户操作过程中,窗口有两种状态,即最大化和自定义大小。当窗口为自定义大小时,可体现用户多窗口协同工作的模式。

5) 窗口切换顺序: 文献[14]观察过计算机工作者经常在并发任务或活动之间切换,并且这种切换行为在工作过程中不断增加。在长期过程中,用户执行活动的顺序具有习惯的模式: 一是用户结合不同功能软件完成一个任务; 二是用户完成多个任务的不同顺序。

6) 窗口切换耗时: 根据文献[15]的研究,用户在不同方向上移动鼠标行为存在显著差异。因此,鼠标不同的移动距离和方向导致点击位置的不同,进而有不同的切换耗时。

7) 窗口打开的星期和时间: 此特征记录用户使用各类型软件在时间上呈现出的周期性。

综上,基于对内网用户的分析研究,结合用户各自的行为度量数据来构建用户的窗口使用行为模式。具体地,每个特征对应的特征值以向量的形式存放,向量内元素的个数对应样本数据的数量,这些特征向量构建成一个特征矩阵空间。将收集到的用户行为数据映射到特征矩阵空间中,对连续特征进行归一化且离散特征进行 One-Hot 编码[16]处理后共得到 130 个连续的特征,最终得到的特征矩阵代表用户的窗口使用行为模式。

1.2　识别用户自身变化的行为特征

如果将变化行为视为异常,变化后的行为中可能仍含有正常的行为模式,例如窗口大小、切换方式和活跃窗口数等。为能够有效识别变化的行为,舍去一些在用户自身行为上差异性小的特征,并将窗口限定为最常用窗口浏览器、Office、论文阅读软件,其余都定义为“其他”,最终特征如下:

1) 窗口偏好: 定义为含有 31 个元素的向量,每个元素对应一个窗口,元素值是使用窗口的经验概率。

2) 单活动持续时长: 定义为含有 4×13 个元素的向量,每个向量记录窗口内单个活动持续时长的经验概率。每 13 个元素的向量初始时长为 0,前 6 个元素宽度设为 20 s,随后 6 个设为 40 s,最后设为无穷大。

3) 活动间隔时长: 定义为含有 4×16 个元素的向量,每个向量记录窗口内活动间隔时长的经验概率。每 16 个元素的向量初始时长为 20,前 11 个元素宽度设为 50 s,随后 4 个设为 100 s,最后设为无穷大。

4) 窗口置顶时长: 定义为含有 4×15 个元素的向量,每个向量记录窗口置顶时长的经验概率。每 15 个元素的向量初始时长为 0,前 10 个元素宽度设为 30 s,随后 4 个设为 60 s,最后设为无穷大。

5) 窗口切换顺序: 定义为含有 31×31 个元素的矩阵。矩阵的每个单元表示窗口之间的切换,其索引由行或列反映。每个单元格的值是窗口切换的概率。

结合用户具体的正常行为数据,给用户建立一个正常的行为模式。首先将终端行为度量值组成 5 个向量元组,并对每个向量进行归一化使向量中所有元素的总和为 1。特别地,活动间隔、持续时长、窗口置顶时长向量乘以 1/4,窗口切换顺序向量乘以 1/31。使用归一化后的元组表示用户正常的行为模式。

1.3　用户行为差异性分析算法

不同用户代表不同的总体,若要准确定义用户使用应用窗口的行为在每个特征向量的分布情况是很困难的,因此借助统计学方法进行差异分析。具

体为:对于特征空间中的每个特征向量借助样本均值的分布来分析用户行为的差异性。

1.3.1 样本均值分布和 K-S 检验

设两个总体对应的分布函数为 $F_1(x)$ 和 $F_2(x)$,由样本均值定理[17]可知,当样本容量 $n(n\geqslant 30)$ 较大时,两个总体对应的样本均值分布为 $\overline{x_1} \dot{\sim} N(\mu_1,\sigma_1^2/n)$ 和 $\overline{x_2} \dot{\sim} N(\mu_2,\sigma_2^2/n)$。因此,若两个样本均值分布不同,则均值或方差也不同,即 $\mu_1\neq\mu_2$ 或 $\sigma_1\neq\sigma_2$,则 $F_1(x)\neq F_2(x)$,表示两个总体的分布也不同。

鉴于上述理论,对每个特征向量构建样本均值分布。采用 K-S 检验[18]对不同用户对应特征的样本均值分布进行差异性检验。K-S 检验方法如下:

设两个用户对应特征的样本均值的累积分布函数为 $F_1(x)$ 和 $F_2(x)$,建立其假设检验,表示为式(1),即两个样本均值的分布情况相同。

$$H_0: F_1(x)=F_2(x) \tag{1}$$

定义检验量为 D,表示为式(2),即两个样本均值的分布情况不同。

$$H_1: F_1(x)\neq F_2(x) \tag{2}$$

$$D=\max\{F_1(x)-F_2(x)\} \tag{3}$$

根据选定的显著性水平 α 值,当 $D>D_{m,n,\alpha}$ 时(m、n 为样本容量,α 为显著性水平),则拒绝假设 H_0,接受 H_1 假设[19]。

实验中从总体中抽样的样本容量为 $n=3\ 000>30$。因此,两个特征对应的样本均值的概率密度函数为:

$$f_1(x)=\frac{1}{\frac{\sigma_1}{n}\sqrt{2\pi}}e^{-(x-\mu_1)^2/2\left(\frac{\sigma_1}{n}\right)^2} \tag{4}$$

$$f_2(x)=\frac{1}{\frac{\sigma_2}{n}\sqrt{2\pi}}e^{-(x-\mu_2)^2/2\left(\frac{\sigma_2}{n}\right)^2} \tag{5}$$

其中,μ_1、μ_2 和 σ_1、σ_2 分别为两个样本总体的均值和方差。由式(3)可知,统计量 D 为两个累积分布函数之差的最大值,即:

$$D=\max\left\{\frac{1}{\frac{\sigma_1}{n}\sqrt{2\pi}}\int_0^x e^{\frac{-(x-\mu_1)^2}{2\left(\frac{\sigma_1}{n}\right)^2}}dx-\frac{1}{\frac{\sigma_2}{n}\sqrt{2\pi}}\int_0^x e^{\frac{-(x-\mu_2)^2}{2\left(\frac{\sigma_2}{n}\right)^2}}dx\right\} \tag{6}$$

统计量 D 对应的显著性水平 p 由可靠性分布函数 $Q_{K\text{-}S}$ 表示:

$$p(D)=Q_{K\text{-}S}(\lambda)=2\sum_{j=1}^{\infty}(-1)^{j-1}e^{-2j^2\lambda^2} \tag{7}$$

其中,$\lambda=\left[\sqrt{N_e}+0.12+\frac{0.11}{\sqrt{N_e}}\right]D$,$N_e=\frac{mn}{m+n}$。若两个独立样本非常相似,则统计量距离 $D\to 0$ 时,$p\to 1$,反之 $p\to 0$。

1.3.2 差异性检验算法

设置拒绝域为 0.05,若显著性水平 $p<0.05$,则认为特征向量的分布来自不同总体,进而说明不同用户的应用窗口使用行为存在显著差异。

算法 1 Difference Verification 算法

输入 待检验的两个用户窗口使用行为数据集 Data1、Data2,显著性水平,样本容量 n,抽样次数 n_{num},特征向量维数 c

输出 1,表示两个样本总体存在显著性差异

1. $[f_1,f_2,\cdots,f_c]$ = FeatureM(Data$_1$)
2. $[g_1,g_2,\cdots,g_c]$ = FeatureM(Data$_2$)
//将用户数据集进行特征工程得到特征矩阵
3. FOR j = 1 to c do
4. μ_{f_j} = Mean(f_j),σ_{f_j} = Var(f_j)
5. μ_{g_j} = Mean(f_j),σ_{g_j} = Var(f_j)
//计算特征矩阵中各特征向量的均值和方差
6. ($X_1,X_2,\cdots,X_{num}$) ←randomSample(Data$_1$,n,n_{num})
($Y_1,Y_2,\cdots,Y_{num}$) ←r andomSample(Data$_2$,n,n_{num})
//抽样 n_{num} 次,生成样本容量为 n 的 n_{num} 个样本空间
7. FOR i = 1 to n_{num} do
8. $[f_1,f_2,\cdots,f_c]_i$←FeatureM(X_i)
9. $[g_1,g_2,\cdots,g_c]_i$←FeatureM(Y_i)
//将每个样本空间进行特征工程生成对应的特征矩阵
10. FOR j = 1 to c do
11. FOR i = 1 to n_{num} do
12. $\bar{f}_{ji}$←Average(f_{ji}),$\bar{g}_{ji}$←Average(g_{ji})
//计算每个特征空间中各个特征向量的均值
13. F($\bar{f}_{ji}$) ←($\bar{f}_{j1},\bar{f}_{j2},\cdots,\bar{f}_{ji}$) $\dot{\sim}$ N($\mu_{f_j},\sigma_{f_j}^2$/n)
G($\bar{g}_{ji}$) ←($\bar{g}_{j1},\bar{g}_{j2},\cdots,\bar{g}_{ji}$) $\dot{\sim}$ N($\mu_{g_j},\sigma_{g_j}^2$/n)
//构建两个用户对应各特征的样本均值分布
14. p(D) = KS(F($\bar{f}_{ji}$),G($\bar{g}_{ji}$))
15. IF p(D) ≤α Then
16. Return True
17. Return 1//K-S 检验两个用户在各个特征上的差异性

在算法 1 中,首先将用户数据集进行特征工程得到特征矩阵,并计算特征矩阵中各特征向量的均值和方差;然后从用户数据集中随机抽样构成 n_{num} 个样本空间并分别进行特征工程,构成 n_{num} 个特征矩阵,计算每个特征矩阵中特征向量的均值,n_{num} 个特征矩阵中对应特征向量的均值共同构成该特征对应的正态分布,且正态分布的均值和方差分别等于该特征向量总体的均值和总体方差的 $1/n$,其中构建各特征的样本均值分布是算法的关键步骤;最后利用 K-S 假设检验验证两个用户对应特征下的样本均值正态分布间的差异,不同用户在相同特征下均表现出差异性便可以说明不同用户行为模式间的差异。

2 用户窗口行为检测

2.1 用户窗口行为检测流程

用户在计算机上的窗口使用行为符合一定的行为模式,依据用户的历史行为数据建立正常的行为模式。根据用户行为模式差异性进行异常检测。具体地,首先将用户的窗口行为数据分成几份,对每一

份分块数据进行特征工程,建立用户的子行为模式;然后计算用户子行为模式之间的距离,依据距离的均值和方差确定用户自身行为波动的置信区间。因为用户的行为模式难以模仿,当一个未知的行为模式与已知用户行为模式之间的距离超出区间的上限时,则判定该未知模式属于异常行为。

2.2　用户窗口行为检测算法

在介绍算法之前,需要先描述量化用户自身行为偏差以及与其他用户的行为差异的公式。

2.2.1　行为偏差量化方法

给定两个表征用户行为的特征矩阵 P 和 Q,每个特征矩阵含有 n 个特征向量。量化行为差异如下:

将 P 中 n 个特征向量与 Q 中对应的向量进行比较。通过计算欧氏距离以量化两个向量之间的差异。给定两个向量 $A=(a_1,a_2,\cdots,a_n)$ 和 $B=(b_1,b_2,\cdots,b_n)$,它们之间的欧氏距离通过式(8)求得。

$$E(A,B)=\sqrt{\sum_{i}^{n}(a_i-b_i)^2} \tag{8}$$

因为相同特征对不同用户的效应不同,利用交叉验证[20]的方法计算特征矩阵中每个特征向量的平均欧氏距离。根据距离均值采用高斯加权法给每个特征赋予权重,特征权重表示为$(w_1,w_2,\cdots,w_n)$。最终利用式(9)计算行为 P 和 Q 的差异为 $\mathrm{WD}(P,Q)$。此值越大,差异越大。

$$\mathrm{WD}(P,Q)=\sqrt{\sum_{j}^{n}\omega_j(E_j)^2} \tag{9}$$

给定用户 self 的行为集合 $\{P_1,P_2,\cdots,P_n\}$,将 self 自身差异定义为每个子集之间的平均差异,具体如下:

$$V_{\mathrm{self}}=\frac{\sum_{i=1}^{n}\sum_{j=1,i\neq j}^{n}\mathrm{WD}(P_i,P_j)}{n(n-1)} \tag{10}$$

利用式(11)计算这些差异的标准偏差:

$$\mathrm{StdDev}=\sqrt{\frac{1}{N}\sum_{i=1}^{n}(\mathrm{WD}(P_i,P_j)-V_{\mathrm{self}})^2} \tag{11}$$

因此,用户 U 的可接受行为偏差以概率 p 落入 $[0,V_{\mathrm{U}}+u\times\mathrm{StdDev}_{\mathrm{U}}]$(假设用户行为符合正态分布)。

2.2.2　异常行为检测算法

异常行为检测算法以行为模式为目标,认为用户自身的行为倾向于遵循其与应用窗口交互的历史行为模式。因为人类行为存在波动性[21],使得一段时间内的行为模式可能会偏移历史行为模式,但是只要发生的偏移在可容忍的波动范围内,则认为当前行为遵循历史行为模式。

对恶意用户而言,其对目标用户的行为习惯知之甚少,因此无法轻易模仿目标用户的行为模式,因此与目标用户自身的波动相比,恶意用户的行为会更加明显地远离目标用户的历史行为模式。将用户自身行为的波动和恶意用户的波动分别称为行为偏差和行为差异。同时,模式间的偏差和差异计算也是异常行为检测算法的关键。

综上,异常行为检测算法由两部分组成: 一是构建用户的子行为模式,计算每个用户自身行为模式偏差的均值、标准差; 二是计算用户自身行为的偏差和与其他用户的行为差异,在保证低假阳性率时确定用户自身行为偏差的置信区间上限,即可容忍的波动范围。

算法 2　SelfDeviation 算法

输入　m 个用户的行为数据$(\mathrm{Data}_1,\mathrm{Data}_2,\cdots,\mathrm{Data}_m)$,每个用户的行为数据含有 N 条窗口活动流,用户数据的分割份数 n,特征向量维数 c

输出　子行为模式、自身行为偏差均值和方差以及权重

1. FOR i = 1 to m do
2. $(\mathrm{Dt}_1,\mathrm{Dt}_2,\cdots,\mathrm{Dt}_n)\leftarrow$noReplaceSample$(\mathrm{Data}_i)$
//不放回抽样 n 次构成 n 个样本空间
3. FOR j = 1 to n do
4. $\mathrm{p}_j=[\mathrm{f}_1,\mathrm{f}_2,\cdots,\mathrm{f}_c]\leftarrow$FeatureM$(\mathrm{Dt}_j)$
5. $\mathrm{P}_i=(\mathrm{p}_1,\mathrm{p}_2,\cdots,\mathrm{p}_n)$
//n 个样本空间进行特征工程形成 n 个特征矩阵
6. FOR j = 1 to n do
7. FOR k = 1, k > j to n do
8. FOR t = 1 to c do
9. $\mathrm{e}_t=$euclideanDist$(\mathrm{P}_i[j][t],\mathrm{P}_i[k][t])$
10. $\mathrm{E}=((\mathrm{e}_{11},\mathrm{e}_{12},\cdots,\mathrm{e}_{1n(n-1)/2}),(\mathrm{e}_{21},\mathrm{e}_{22},\cdots,\mathrm{e}_{2n(n-1)/2}),\cdots,(\mathrm{e}_{c1},\mathrm{e}_{c2},\cdots,\mathrm{e}_{cn(n-1)/2}))$
//计算 n 个特征矩阵中对应特征向量间的欧氏距离
11. $\bar{\mathrm{E}}=(\bar{\mathrm{e}}_1,\bar{\mathrm{e}}_2,\cdots,\bar{\mathrm{e}}_c)\leftarrow$averageDist$(\mathrm{E})$
//计算各个特征的平均欧式距离
12. $\mathrm{W}_i=(\mathrm{w}_1,\mathrm{w}_2,\cdots,\mathrm{w}_c)\leftarrow$GaussionWeight$(\bar{\mathrm{E}})$
//依据平均欧氏距离求特征对应的权重
13. FOR j = 1 to n do
14. FOR k = 1, k > j to n do
15. $\mathrm{we}_{jk}\leftarrow$euclideanDist$(\mathrm{W}_i,\mathrm{P}_i[j],\mathrm{P}_i[k])$
16. $\mathrm{WE}_i=(\mathrm{we}_{11},\mathrm{we}_{12},\cdots,\mathrm{we}_{(n-1)n})$
//计算各个子模式之间的加权欧氏距离
17. $\mathrm{v}_i=$averageDiffence(WE_i), $\mathrm{sd}_i=$StdDev(WE_i)
//计算各子模式间距离的平均值和标准偏差
18. $\mathrm{U}=((\mathrm{p}_{11},\mathrm{p}_{12},\cdots,\mathrm{p}_{1n}),(\mathrm{p}_{21},\mathrm{p}_{22},\cdots,\mathrm{p}_{2n}),\cdots,(\mathrm{p}_{m1},\mathrm{p}_{m2},\cdots,\mathrm{p}_{mn}))$, $\mathrm{W}=(\mathrm{W}_1,\mathrm{W}_2,\cdots,\mathrm{W}_m)$, $\mathrm{V}=(\mathrm{v}_1,\mathrm{v}_2,\cdots,\mathrm{v}_m)$, $\mathrm{StdDev}=(\mathrm{sd}_1,\mathrm{sd}_2,\cdots,\mathrm{sd}_m)$
19. Return U, W, V, StdDev
//返回所有用户的子行为模式、特征权重、用户自身行为
//偏差的均值和方差

在算法 2 中,首先对用户的历史行为数据进行不放回抽样构成样本空间,将样本空间分别进行特征工程得到用户子特征矩阵代表的子行为模式,利用交叉验证方法计算子行为特征矩阵内对应特征向量之间的平均欧式距离,依据平均欧式距离计算每个特征向量的权重; 然后计算每个子模式之间的加权欧氏距离,并计算距离的均值和标准偏差,最终得到用户 U 的可接受行为偏差以概率 p 落入区间 $[0,V_{\mathrm{U}}+u\times\mathrm{StdDev}_{\mathrm{U}}]$ 的参数 V_{U} 和 $\mathrm{StdDev}_{\mathrm{U}}$。

算法 3　DetermineConfidenceInterval 算法

输入　m 个用户的行为模式 BP,每个用户的对应的特征权重 W,用户行为偏差的均值 V 和标准差 SD,标准差的倍数

$u=(1.31,1.32,\cdots,2.35)$

输出 用户自身行为偏差的置信区间上限

```
FOR i = 1 to m do
FOR j = 1, j≠i to m do
FOR k = 1 to n do
we_ij ← euclideanDist(W_i, BP[i][k], BP[j][k])
WE_i = (we_1, we_2, ···, we_(m-1)n^2)
//计算用户自身与其他用户的加权欧式距离
num_i = length(BP[i]), num_j = length(BP[j])
//用户自身子模式数,其他用户子行为模式总数
FOR v in u do
threshold = V[i] + v × SD[i]
FOR we in WE_i do
num = count(we > threshold)
FP = num ÷ (m - 1) × num_i × num_j //计算假阳性率
IF FP < 4% Then
best = v
bestThreshold = V[i] + best × SD[i]
//在低假阳性率下确定用户自身行为波动阈值
Return m_bestThreshold
```

在算法 3 中,首先计算不同用户行为模式之间的差异,并和用户自身行为偏差进行对比,在保证低假阳性率的条件下求得自身行为波动的置信区间,即确定区间$[0,V_U+u\times \mathrm{StdDev}_U]$中的参数 u。

综上,当输入新的未知行为数据后,计算与已知用户的正常的行为模式之间的差异距离 WE,如果$\mathrm{WE}>V_U+u\times \mathrm{StdDev}_U$,则说明当前输入的未知行为数据为非法数据,否则为合法数据。

3 实验结果与分析

3.1 实验环境与数据

为更加全面地描述用户使用窗口的行为模式,通过详细分析用户与计算机窗口交互时最直接的操作特性,依据交互特性提出行为特征,并以特征为目标开发数据采集器。其优势是分析用户行为时不会因受制于系统数据的单一性,而不能全面地建立用户的行为模式,且避免了从复杂系统数据中挖掘用户数据时带来的巨大工作量。

为获取真实内网用户数据,本文搭建了完全自由的内网环境,采集器使用键盘鼠标触发模式,在数据录入过程中不做处理,避免了数据采集延迟。对 20 位用户收集两个月的窗口活动数据。手动检查第一周数据后,最终筛选出 15 位用户的数据。数据以“窗口名”为单位进行组织,当用户在一个窗口中使用鼠标或键盘时就表示用户与窗口开始交互。当鼠标键盘动作切换到另一窗口时,就表明与前一个窗口结束交互且与下一个窗口交互开始。采集阶段共收集了 30 多万条交互数据。

3.2 差异性分析

3.2.1 用户间行为差异和自身行为偏差

将 15 位用户的窗口行为数据集依据不放回抽样分为 5 个样本容量为 3 000 的样本空间,经特征工程后构成子行为模式。利用算法 2 分别计算用户自身子行为模式间和用户与其他用户子行为模式间距离的平均值与标准偏差,用以表示行为偏差和行为差异。实验结果证明,相比用户自身行为偏差,用户间的行为差异更为明显。

图 3 显示了每个用户自身的行为偏差和与其他用户的平均行为差异。由图 3 可知,每个用户自身偏差明显低于与其他用户的平均行为差异,相比用户间的行为差异,用户自身的行为偏差通常在一个小范围内波动。此结果也表明利用用户行为模式间的差异性进行异常检测是可行的。

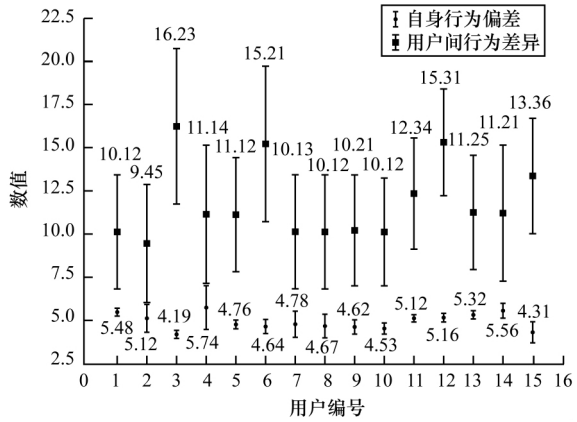

图 3 自身行为偏差和用户间行为差异

Fig. 3 Self-behavioral bias and behavioral differences between users

任选两个用户的窗口行为数据集,设定检验的显著性水平 $\alpha=0.05$,样本容量 $n=3\ 000$,有放回抽样次数 $n_{\mathrm{num}}=30$。利用差异性检验算法(算法 1)检验用户窗口行为在各个特征上呈现的差异性。检验结果如表 1 所示。

表 1 用户行为在特征向量上的差异性检验结果

Table 1 Difference test results of user behaviors on feature vectors

特征	统计量 D	P 值	特征差异是否显著
置顶时长	0.766 6	0.000 0	1
单个活动均值	0.366 6	0.025 5	1
有效总时长	0.433 3	0.004 6	1
间隔时长均值	0.533 4	0.000 2	1
间隔时长方差	0.366 6	0.025 8	1
⋮	⋮	⋮	⋮
间隔次数	0.533 3	0.000 2	1
鼠标动作时长	0.600 0	0.000 0	1
键盘动作时长	0.533 3	0.000 2	1
自定义窗口大小	0.433 3	0.004 6	1
活跃窗口个数	0.233 3	0.006 0	1

在表 1 中,第 4 列中的结果值为“1”表明用户的行为数据在各个特征上表现出显著的差异性,说明不同用户同一特征对应的样本均值分布的差异显著,依据样本均值定理该结果也表明不同用户的行为模式之间存在显著的差异性。

3.2.2　用户行为变化前后的差异

依据无放回抽样分别从用户 1 万条变化后的数据中进行 6 次抽样和正常数据中进行 10 次抽样构成样本容量为 1 500 的 6 个负样本空间和 10 个正样本空间,利用算法 2 计算正样本各子模式之间距离的平均值与标准偏差和负样本子模式与正样本子模式之间距离的平均值与标准偏差。

图 4 显示了用户行为变化前后的行为差异。由此可知,当用户行为发生变化后,变化的行为模式明显远离正常行为的波动阈值。不同变化程度的行为模式与波动阈值的距离不同。当用户行为波动范围超过阈值后将其判为变化的行为,由此对用户变化的行为进行监控。

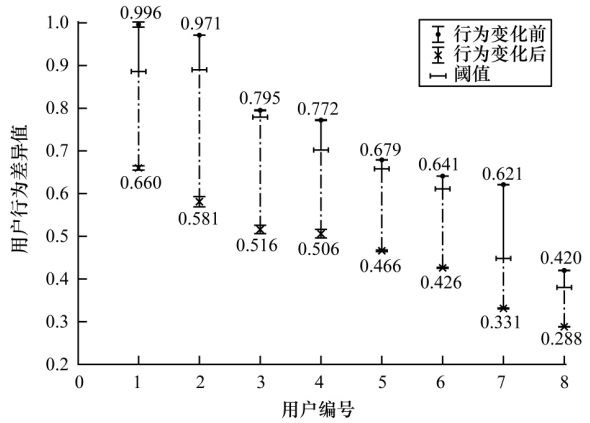

图 4　用户行为变化前后的行为差异

Fig. 4　Behavior difference before and after the change of user behavior

3.3　性能评估

由异常行为检测算法可知,区间上限 $V_U + u \times \mathrm{StdDev}_U$ 中的 u 取值越大对应用户行为波动偏差的区间上限越大,待检测样本落入正常波动区间的概率也增大,即判定待检测的用户行为不属于异常行为的概率更大。因此,异常检测准确率和 u 是线性关系。但当波动偏差的区间上限过度增大时,容易将异常行为判定为正常,导致很高的假阳性率。因此,将准确率和假阳性率作为评价指标进行评估。

3.3.1　异常用户检测

将所有用户的数据集进行 5 次不放回抽样得到每个用户的 5 个子行为模式,其中,用户自身的子行为模式作为正样本,其他用户的所有子行为模式作为负样本。在训练过程中,选取用户自身的全部 5 个子行为模式,从其他用户子行为模式中分别选取 2 个子行为模式,组成 14×2 个负样本子行为模式,然后利用异常检测算法在保证低假阳性率的情况下确定用户行为波动阈值上限 $V_U + u \times \mathrm{StdDev}_U$ 中的 u 值。在验证过程中,将其余的 14×3 个子模式作为验证集,计算各个负样本子模式与正样本子模式之间的距离,并计算 u 值对应的假阳性率。为使每个子模式都参与训练和验证,且结果不是来自偏差数据,利用交叉验证方法进行 10 次实验。假阳性率为 $F_{FP} = \frac{n'_{num}}{(m-1) \times n_1 \times n_2}$,其中,$n'_{num}$ 是将异常模式判为正常模式的个数,用户数 m 为 15,n_1、n_2 分别为正常用户和异常用户的子模式个数,分别为 5 和 3。

图 5 显示了当 u 的取值在 2.0 附近时,假阳性率增加,但准确率不会明显提高。因此,在保证低假阳性率(4%)时 u 的取值为 1.95,对应的准确率为 97.4%,即当前检测的行为模式为异常模式的概率为 97.4%。

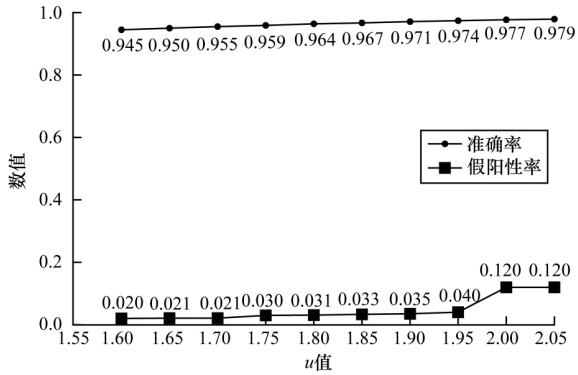

图 5　u 值与准确率和假阳性率的关系

Fig. 5　Relationship between the value of u, accuracy and the false positive rate

构建行为模式的样本数据越多可以更好地反映用户行为。在 u 取值为 1.95 时,分别使用 500,750,…,3 000 条活动流构建用户的子行为模式。计算不同数量的活动流构建的行为模式对应的假阳性率。

图 6 显示了随着训练集大小的变化,假阳性率的动态变化。总的来看,训练集越大,检测的假阳性率越低,从而反映活动流越长,提供的用户习惯行为数据越多,使得能够建立更完整的行为模式,因此,可以更准确地测量用户在每个行为特征上的距离,进而使得检测结果更准确。

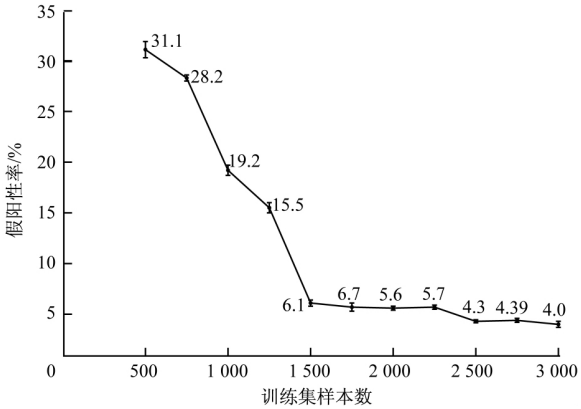

图 6　训练集大小与假阳性率的关系

Fig. 6　Relationship between the size of the training set and the false positive rate

3.3.2　用户变化行为识别

在数据采集周期内,有 8 个用户因工作进展其阶段性任务发生了变化,阶段任务变化的事件如表 2所示。

表 2 影响用户行为的事件

Table 2 Events that affect user behavior

用户	阶段任务变化的事件
User 1	论文投稿
User 2	课程结束
User 3	课程结束
User 4	实验室网站搭建
User 5	项目结题
User 6	云服务器诚信实验
User 7	大数据平台搭建
User 8	博士考试

以事件前后的数据为正负样本进行实验。依据无放回抽样分别从用户 1 万条变化后的数据中进行 6 次抽样和正常数据中进行 10 次抽样构成样本容量为 1 500 的 6 个负样本空间和 10 个正样本空间，经特征工程后产生对应的子行为模式。在训练过程中，选取 3 个负样本子行为模式和 10 个正样本子行为模式。验证过程中计算另外 3 个负子行为模式与正样本子行为模式之间的距离。其中，假阳性率 $F_{FP}=\frac{n_{误判}}{n_{正}\times n_{负}}$，$n_{误判}$是将异常模式判为正常模式的个数，$n_{正}$和 $n_{负}$分别是正常和异常子模式的个数。利用异常检测算法对变化的行为进行识别，在保证较低的假阳性率(9%)时 u 的取值为 1.6，对应的检测准确率为 94.5%。

对于用户变化行为的识别而言，用户的行为模式存在波动是正常的，但当波动超过可容忍的范围后就认为当前的行为发生了变化。此时可对用户发生变化的行为提出预警，然后结合用户背景详细分析用户的操作，确定用户变化的行为是否为恶意操作，若是则可及时止损。

4 结束语

本文从计算机应用窗口的角度出发对用户行为进行研究。针对用户分类和用户变化行为的识别分别提出一系列可以表征用户行为的特征。在完全自由的内网环境中，采集用户使用应用窗口的行为数据。通过样本均值特性和 K-S 检验，验证了用户使用应用窗口行为存在显著的差异性，并表明本文构建的用户行为检测算法能够有效检测异常用户和识别用户变化的行为。后续将尝试在更大的内网环境中，以角色为团体进行窗口行为的异常检测研究。

参考文献

[1] LANE T, BRODLEY C E. An empirical study of two approaches to sequence learning for anomaly detection [J]. Machine Learning, 2003, 51(1): 73-107.

[2] YANG Guang, MA Jiangang, YU Aimin, et al. Survey of insider threat detection [J]. Journal of Cyber Security, 2016, 1(3): 21-36. (in Chinese)
杨光，马建刚，于爱民，等. 内部威胁检测研究 [J]. 信息安全学报，2016，1(3)：21-36.

[3] LU Jun, LIU Daxin, FU Liping. Research and design of dynamic security model for preventing network internal threats [J]. Computer System Applications, 2005, 14(9): 37-40. (in Chinese)
陆军，刘大昕，付立平. 防范网络内部威胁的动态安全模型的研究与设计 [J]. 计算机系统应用，2005，14(9)：37-40.

[4] LI Dianwei, HE Mingliang, YUAN Fang. Research on insider threat detection based on role behavior pattern mining [J]. Information Network Security, 2017, 17(3): 27-32. (in Chinese)
李殿伟，何明亮，袁方. 基于角色行为模式挖掘的内部威胁检测研究 [J]. 信息网络安全，2017，17(3)：27-32.

[5] OKA M, OYAMA Y, ABE H. Anomaly detection using layered networks based on eigen co-occurrence matrix [C]// Proceedings of the 7th International Symposium on Recent Advances in Intrusion Detection. Sophia Antipolis, France: [s.n.], 2004: 223-237.

[6] SCHONLAU M, DUMOUCHEL W, JU W H, et al. Computer intrusion: detecting masquerades [J]. Statistical Science, 2001, 16(1): 58-74.

[7] NGUYEN N, REIHER P, KUENNING G H. Detecting insider threats by monitoring system call activity [C]// Proceedings of Information Assurance Workshop. Washington D.C., USA: IEEE Press, 2004: 3-10.

[8] HUANG Tie, ZHANG Fen. Hidden Markov model for internal threat detection [J]. Computer Engineering and Design, 2010, 31(5): 965-968. (in Chinese)
黄铁，张奋. 基于隐马尔可夫模型的内部威胁检测方法 [J]. 计算机工程与设计，2010，31(5)：965-968.

[9] LI L, MANIKOPOULOS C N. Windows NT one-class masquerade detection [C]//Proceedings of the 5th Annual IEEE SMC Information Assurance Workshop. Washington D.C., USA: IEEE Press, 2004: 447-465.

[10] ZHANG Rui, CHEN Xiaojun, SHI Jinqiao, et al. Detecting insider threat based on document access behavior analysis [C]//Proceedings of Asia Pacific Web Conference. Berlin, Germany: Springer, 2014: 376-387.

[11] CAMI J B, RODR J. Towards a masquerade detection system based on user's tasks [C]//Proceedings of the 17th International Workshop on Recent Advances in Intrusion Detection. Berlin, Germany: Springer, 2014: 447-465.

[12] GOLDRING T. Authenticating users by profiling behavior [EB/OL]. [2019-07-05]. https://cs.fit.edu/~pkc/dmsec03/slides.

[13] WEI Bocheng, LIN Jinguan, XIE Fengchang. Statistical diagnosis [M]. Beijing: Higher Education Press, 2009. (in Chinese)
韦博成，林金官，解锋昌. 统计诊断 [M]. 北京：高等教育出版社，2009.

(下转第 150 页)

[6] HUANG Kun,XIAN Ming,FU Shaojing,et al. Securing the cloud storage audit service: defending against frame and collude attacks of third party auditor [J]. IET Communications,2014,8(12):2106-2113.

[7] WANG Zhihao. Cloud storage data integrity verification scheme and its improvement [D]. Chengdu: Southwest Jiaotong University,2018. (in Chinese)
王志豪. 云存储数据完整性验证方案及其改进 [D]. 成都: 西南交通大学,2018.

[8] ZHOU Enguang, LI Zhoujun, GUO Hua, et al. An improved data integrity verification scheme in cloud storage system [J]. Acta Electronica Sinica, 2014, 42(1):150-154. (in Chinese)
周恩光,李舟军,郭华,等. 一个改进的云存储数据完整性验证方案 [J]. 电子学报,2014,42(1):150-154.

[9] ZHA Yaxing,LUO Shoushan,LI Wei,et al. Dynamic group public auditing scheme for shared data on attribute-based threshold signature [J]. Journal of Beijing University of Posts and Telecommunications,2017,40(5):43-49. (in Chinese)
查雅行,罗守山,李伟,等. 基于属性门限签名的动态群组共享数据公开审计方案 [J]. 北京邮电大学学报,2017,40(5):43-49.

[10] HUANG Longxia, ZHANG Gongxuan, FU Anmin. Privacy-preserving public auditing for dynamic group based on hierarchical tree [J]. Journal of Computer Research and Development,2016,53(10):2334-2342. (in Chinese)
黄龙霞,张功萱,付安民. 基于层次树的动态群组隐私保护公开审计方案 [J]. 计算机研究与发展,2016,53(10):2334-2342.

[11] FU Anmin,QIN Ningyuan,SONG Jianye,et al. Privacy-preserving public auditing for multiple managers shared data in the cloud [J]. Journal of Computer Research and Development,2015,52(10):2353-2362. (in Chinese)
付安民,秦宁元,宋建业,等. 云端多管理者群组共享数据中具有隐私保护的公开审计方案 [J]. 计算机研究与发展,2015,52(10):2353-2362.

[12] DOMINGO F J, QIN B, WU Q, et al. Identity-based remote data possession checking in public clouds [J]. IET Information Security,2014,8(2):114-121.

[13] TAN Shuang,JIA Yan. NaEPASC: a novel and efficient public auditing scheme for cloud data [J]. Journal of Zhejiang University-Science C,2014,15(9):794-804.

[14] WANG Huaqun. Identity-based distributed provable data possession in multicloud storage [J]. IEEE Transactions on Services Computing,2015,8(2):328-340.

[15] HUANG Longxia, ZHANG Gongxuan, FU Anmin. Privacy-preserving public auditing for non-manager group [C]// Proceedings of 2017 IEEE International Conference on Communications. Washington D.C., USA: IEEE Press,2017:1-6.

[16] LIU Hongyu,MU Yi,ZHAO Jining,et al. Identity-based provable data possession revisited: security analysis and generic construction [J]. Computer Standards and Interfaces, 2017,54(1):10-19.

[17] ALRIYAMI S S,PATERSON K G. Certificateless public key cryptography [C]// Proceedings of ASIACRYPT'03. Berlin, Germany: Springer,2003:452-473.

[18] WANG Boyang,LI Baochun,LI Hui,et al. Certificateless public auditing for data integrity in the cloud [C]// Proceedings of 2013 IEEE Conference on Communications and Network Security. Washington D.C., USA: IEEE Press,2013:136-144.

[19] WU L,JING W,ZEADALLY S,et al. Privacy-preserving auditing scheme for shared data in public clouds [J]. Journal of Supercomputing,2018,74(11):6156-6183.

[20] WANG Boyang,LI Baochun,LI Hui. Panda: public auditing for shared data with efficient user revocation in the cloud [J]. IEEE Transactions on Services Computing,2015,8(1):92-106.

编辑 司淼淼

(上接第 142 页)

[14] BANNON L,CYPHER A,GREENSPAN S,et al. Evaluation and analysis of users' activity organization [C]// Proceedings of ACM SIGCHI Conference on Human Factors in Computing Systems. New York,USA: ACM Press,1983:54-57.

[15] AHMED A A E,TRAORE I. A new biometric technology based on mouse dynamics [J]. IEEE Transactions on Dependable and Secure Computing,2007,4(3):165-179.

[16] CHEN Guangxin. Proficient in feature engineering [M]. Beijing: People's Posts and Telecom Press,2019. (in Chinese)
陈光欣. 精通特征工程 [M]. 北京: 人民邮电出版社,2019.

[17] SHI Shisong, CHENG Yiming, XIAO Xiaolong. Probability theory and mathematical statistics [M]. Beijing: Higher Education Press,2011. (in Chinese)
茆诗松,程依明,濮晓龙. 概率论与数理统计教程 [M]. 北京: 高等教育出版社,2011.

[18] WU Xizhi,WANG Zhaojun. Nonparametric statistical method [M]. Beijing: Higher Education Press,1996. (in Chinese)
吴喜之,王兆军. 非参数统计方法 [M]. 北京: 高等教育出版社,1996.

[19] WU Xizhi. Statistics: from data to conclusion [M]. Beijing: China Statistics Press,2014. (in Chinese)
吴喜之. 统计学: 从数据到结论 [M]. 北京: 中国统计出版社,2014.

[20] ZHOU Zhihua. Machine learning [M]. Beijing: Tsinghua University Press,2016. (in Chinese)
周志华. 机器学习 [M]. 北京: 清华大学出版社,2016.

[21] ZHOU Tao,HAN Xiaoyu,YAN Xiaoyong,et al. Statistical mechanics on temporal and spatial activities of human [J]. Journal of University of Electronic Science and Technology of China,2013,42(4):481-540. (in Chinese)
周涛,韩筱璞,闫小勇,等. 人类行为时空特性的统计力学 [J]. 电子科技大学学报,2013,42(4):481-540.

编辑 陆燕菲