基于用户鼠标行为的身份认证方法

徐 剑[1,2] 李明洁[1] 周福才[1] 薛 锐[2]
(东北大学软件学院 沈阳 110819)[1]
(中国科学院信息工程研究所信息安全国家重点实验室 北京 100093)[2]

摘 要 针对已有身份认证方法存在的问题,提出了一种基于用户鼠标行为的身份认证方法。首先,给出了基于用户鼠标行为的身份认证模型及其实体构成,之后采用层次化划分方法对用户鼠标行为进行了定义,同时给出了不同行为需要提取的特征值及对应的计算方法。采用随机森林分类器作为鼠标行为分类工具,以解决已有方案中存在的数据的过度拟合和数据噪声问题。在身份认证阶段,结合用户鼠标行为,采用层次结构的分类决策模型对用户身份进行认证。最后,对所提出的身份认证方法进行了实验分析,结果表明该方法具有较好的错误拒绝率和错误接受率。

关键词 身份认证,鼠标行为,随机森林
中图法分类号 TP309 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2016.2.033

Identity Authentication Method Based on User's Mouse Behavior

XU Jian[1,2] LI Ming-jie[1] ZHOU Fu-cai[1] XUE Rui[2]
(Software College, Northeastern University, Shenyang 110819, China)[1]
(State Key Laboratory of Information Security, Institute of Information Engineering, Chinese Academy of Sciences, Beijing 100093, China)[2]

Abstract Aiming at the problems of existing solution for identity authentication, the paper provided a method about the identity authentication based on mouse behaviour. First, it gave the authentication model based on the mouse action and the entity set, and then defined the mouse action by hierarchical division method. At the same time, it gave the definition of features related to the mouse action and the computing method. The paper used random forest classifier as the classification tool to solve the problem of data over-fitting and noise in the existing solution. In the phase of authentication, it used hierarchical classification-decision model for identity authentication. Finally, the paper analyzed the method through experiment, showing a better false rejection rate and false acceptance rate.

Keywords Identity authentication, Mouse behavior, Random forest

1 引言

用户身份认证是计算机及网络系统的重要安全保障手段之一。现有的身份认证技术主要有3类:口令认证、持有物认证和生物认证。口令认证存在容易被泄露、破解的风险;持有物认证则存在持有物易丢失或被伪造的缺陷。基于生物特征的用户身份认证与口令认证和持有物认证相比,在安全性上有了很大的改进,避免了认证信息泄露、破解、丢失以及仿冒等问题。

目前,基于生物特征的身份认证方法主要有两类,一类是基于用户生理特征的身份认证,另一类是基于用户行为特征的身份认证。基于用户生理特征的身份认证方法中常用的识别技术主要有指纹识别[1]、掌纹识别[2]、人脸识别[3]以及虹膜识别[4]等。这些识别技术不仅需要额外的硬件识别设备,而且很有可能泄露用户的隐私信息。因此,基于用户生理特征的身份认证方法并不适用于当前的互联网环境。而基于用户行为特征的身份认证方法利用人机交互设备对用户行为数据进行采集,不需要额外的硬件设备,同时,数据采集过程并不影响用户的正常操作,且可以在当前多数计算机系统中直接部署,因此逐渐成为当前身份认证研究领域中的热点内容之一。

在目前的计算机和网络应用环境中,键盘和鼠标是应用最广泛的人机交互设备。目前已有很多学者对基于用户击键行为的身份认证技术进行了较为深入的研究[5,6]。但随着图形化界面的日益普及,鼠标已逐渐超越键盘成为当前人机交互环境中的主要输入设备,并受到越来越多研究者的关注。

2003年,Ahmed等人[7]首次证明了利用鼠标行为特征对用户进行身份认证的可行性,对用户鼠标行为中鼠标移动速度、鼠标移动距离、单击次数等物理量进行统计分析,结果表明在不同的用户间,这些统计量存在差异,并提出基于这些差异识别用户身份的初步方案。随后,很多学者陆续开展了基于用户鼠标行为的身份认证研究。目前基于鼠标行为的身

到稿日期:2015-02-26 返修日期:2015-05-25 本文受国家科技重大专项基金项目(2013ZX03002006),辽宁省科技攻关项目(2013217004),辽宁省博士启动基金项目(20141012),中央高校基本科研业务费专项资金(N130317002),信息安全国家重点实验室开放课题项目(2014-15)资助。
徐 剑(1978-),男,博士,讲师,主要研究方向为密码学与网络安全、云计算安全等;李明洁(1992-),女,硕士生,主要研究方向为云计算安全;周福才(1964-),男,教授,博士生导师,主要研究方向为密码学与网络安全、可信计算等;薛 锐(1963-),男,研究员,博士生导师,主要研究方向为密码学与网络安全、安全协议的设计与分析。

份认证方案根据认证模式的不同可以分为两类[8]，即静态认证和持续认证。静态认证是指要求用户在一个特定的时间内完成某项预先设定好的任务，在这一过程中对用户鼠标数据进行采集，并进行身份认证。而持续认证方法是指用户行为的采集和认证操作贯穿于用户与计算设备交互的整个过程。Gamboa 等人[9]要求用户使用鼠标在屏幕上的虚拟键盘上输入自己的标识码，进而达到识别用户身份的目的，该方法就是一种静态认证方法。Pusara 和 Bordley[10]利用用户浏览网页时的鼠标行为对用户进行身份认证，该方法是一种持续认证方法。Bours 等人[11]提出一种静态的认证方法，当用户使用鼠标穿越一个屏幕迷宫时记录用户的鼠标活动，提取这一过程中的速度向量，通过比较两个速度向量的相似度来对用户的身份进行认证。Schulz 等人[12]提出了一种切割鼠标手势的方法，将鼠标手势切割成一段一段的具有长度、曲率等特征的曲线，然后对这些曲线提取特征值，构成特征直方图，在认证阶段，通过比较直方图来进行身份认证，该方法也是一种持续认证方法。Ahmed 等人[13]提出的持续认证方法是在用户日常工作中收集数据，对数据进行特征值提取，然后将特征值聚集成直方图形式，每个用户的行为特征使用 7 个直方图来标记，使用人工神经网络的方法来进行比较分析，从而进行身份认证。沈超等人[14]提出了一种利用人机交互时计算机用户的鼠标使用行为特征进行身份认证和监控的方法，该方法也是一种持续认证方法。

基于鼠标行为的身份认证方法的研究还处于起步阶段。研究者们为了提高鼠标行为身份认证的准确性和可行性，对要采集的鼠标行为和使用的鼠标特征值给出了不同的实验定义，对采集得到的鼠标特征值的处理也不相同。但是多数已有方法采用的是基于单决策树的分类器构建方法，因此，存在数据过度拟合和噪声影响的问题。为此，本文首先给出了基于用户鼠标行为的身份认证模型，并采用层次化划分的方法对鼠标行为进行定义，同时给出了不同行为需要提取的特征值。在身份认证阶段，结合用户鼠标行为，采用层次结构的分类决策模型对用户身份进行认证。采用随机森林分类器作为鼠标行为分类的工具，该分类器同其他基于单决策树的分类器相比，可以避免数据的过度拟合和数据噪声的影响。最后，对本文提出的方法进行实验分析，结果表明本方法具有较好的错误拒绝率和错误接受率。

2 基于用户鼠标行为的身份认证模型

2.1 模型框架

基于用户鼠标行为的身份认证模型主要包括 4 个部分：鼠标行为预处理、鼠标行为特征值提取、分类模型的构建以及分类决策，如图 1 所示。

(1)鼠标行为预处理过程包括基本鼠标事件的采集和鼠标行为的组合，在这一过程中通过一定规则将捕获到的鼠标基本事件组合成可用于特征值计算的鼠标行为。

(2)鼠标特征值提取是对捕获到的不同鼠标类型的行为提取其相应的特征值，这些特征值经过离散化处理后将用来进行分类模型的训练，或者作为分类决策模型的输入对用户身份进行认证。

(3)训练建模使用采集到的不同用户的行为特征值来构建一个分类模型，该模型对监控到的一组鼠标特征值数据进行所属关系的分类预测。

(4)分类决策是通过监控到的鼠标特征值数据对用户身份进行认证，该过程分为两个层：分类层和决策层。

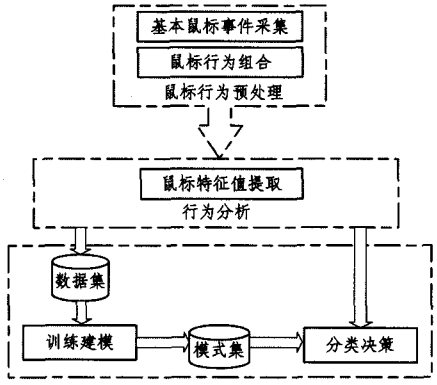

图 1 基于用户鼠标行为的身份认证模型

2.2 鼠标事件及行为定义

为了有效描述用户鼠标行为，对鼠标行为进行层次化的划分。所有的高级鼠标行为由低级的鼠标行为按照一定的规则组合而成，所有的鼠标行为都能拆分成独立的基本鼠标事件序列。

2.2.1 鼠标事件

所有的鼠标行为都能拆分成基本鼠标事件组成的序列，下面给出基本鼠标事件的相关描述。

(1)鼠标移动事件(m)：当鼠标从一点移动到另一点时触发该事件。

(2)鼠标按下事件(down)：鼠标按键被按下时触发该事件，本文只考虑鼠标左键按下(ld)和右键按下(rd)。

(3)鼠标弹起事件(up)：鼠标按键弹起时触发该事件，本文只考虑鼠标左键弹起(lu)和右键弹起(ru)。

本文中的鼠标事件以四元组的形式〈type, time, (x, y)〉来表示。type 表示类型，time 表示捕获的时间，x 表示鼠标事件捕获位置的横坐标，y 表示鼠标事件捕获位置的纵坐标。

2.2.2 鼠标行为

本文将鼠标行为划分为 3 个层次，低层次(层次 I)的鼠标行为由一系列的基本鼠标事件(包括鼠标移动、按下以及弹起)构成，高层次(层次 II 和 III)的鼠标行为由低层次的鼠标行为按照一定的规则链接而成。

(1)第 I 层的鼠标行为

第 I 层的鼠标行为包括鼠标的左键单击(LC)、右键单击(RC)、左键拖拽(LDD)、右键拖拽(RDD)以及鼠标移动序列(MMS)。

为了将基本的鼠标事件进行组合，定义 1 对第 I 层中的时间阈值的概念进行了描述。

定义 1 第 I 层中的时间阈值

①移动事件阈值 τ_{mm}：将两个移动事件连接在一起形成移动序列的最大时间间隔。

②单击行为时间阈值：形成鼠标单击行为的鼠标按下和鼠标弹起之间的最大时间间隔，包括左键单击阈值(τ_{LC})和右键单击阈值(τ_{RC})。

③拖拽行为时间阈值：鼠标按键按下和鼠标按键弹起能组成一个行为的最小时间间隔，本文包括左键拖拽阈值(τ_{LDD})和右键拖拽阈值(τ_{RDD})。

在定义 1 的基础中，给出第 I 层鼠标行为的相关定义，如定义 2 所描述。

定义 2 第 I 层中的鼠标行为定义

①鼠标左键单击(LC),是指敲击鼠标左键的行为,该行为在 τ_{LC}(按下鼠标左键到弹起鼠标左键的时间间隔)内发生,如式(1)所示。

$$LC_{t_1}^{t_n}=\langle ld_{t_1},[m_{t_2},m_{t_3},\cdots,m_{t_{n-1}}],lu_{t_n}\mid t_n-t_1\leqslant\tau_{LC}\rangle \quad (1)$$

②鼠标右键单击(RC),是指敲击鼠标右键的行为,该行为在 τ_{RC}(按下鼠标右键到弹起鼠标右键的时间间隔)内发生,如式(2)所示。

$$RC_{t_1}^{t_n}=\langle rd_{t_1},[m_{t_2},m_{t_3},\cdots,m_{t_{n-1}}],ru_{t_n}\mid t_n-t_1\leqslant\tau_{RC}\rangle \quad (2)$$

③鼠标左键拖拽行为(LDD),指从按下鼠标左键开始,移动一段时间后,以一次鼠标左键弹起结束,左键按下与左键弹起的时间间隔大于 τ_{LDD},如式(3)所示。

$$LDD_{t_1}^{t_n}=\langle ld_{t_1},[m_{t_2},m_{t_3},\cdots,m_{t_{n-1}}],lu_{t_n}\mid t_n-t_1>\tau_{LDD}\rangle \quad (3)$$

④鼠标右键拖拽行为(RDD),指从按下鼠标右键开始,移动一段时间后,以一次鼠标右键弹起结束,左键按下与左键弹起的时间间隔大于 τ_{RDD},如式(4)所示。

$$RDD_{t_1}^{t_n}=\langle rd_{t_1},[m_{t_2},m_{t_3},\cdots,m_{t_{n-1}}],ru_{t_n}\mid t_n-t_1>\tau_{RDD}\rangle \quad (4)$$

⑤鼠标移动序列(MMS),是指一个鼠标移动事件序列,序列内两个按时间顺序发生的鼠标移动事件之间的时间间隔小于 τ_{mm},如式(5)所示。

$$MMS_{t_1}^{t_n}=\langle m_{t_1},m_{t_2},\cdots,m_{t_n}\mid\forall k:1\leqslant k\leqslant n-1,t_{k+1}-t_k\leqslant\tau_{MM}\rangle \quad (5)$$

(2)第 II 层的鼠标行为

第 II 层的鼠标行为包括鼠标左键双击(DC)和鼠标移动(MM)。该层中的时间阈值定义将在定义 3 中给出。

定义 3 第 II 层中的时间阈值

①静态间隔(σ),在两个按时间顺序发生的鼠标事件之间没有行为发生,这期间的时间间隔定义为静态间隔。

②双击时间阈值(τ_I),两次按时间顺序发生的鼠标单击行为能够连接成一次鼠标双击行为之间的最大时间间隔。

在定义 3 的基础上给出第 II 层鼠标行为的定义,如定义 4 所示。

定义 4 第 II 层中的鼠标行为定义

①鼠标移动(MM),由一个鼠标移动序列和一段时间的静态间隔(σ)组成,如式(6)所示。

$$MM=(MMS,\sigma) \quad (6)$$

②鼠标双击(DC),指鼠标左键双击行为,由两个连续的鼠标左键单击行为组成,第一次鼠标单击行为中左键弹起的时间和第二次鼠标单击行为中左键按下的时间之间的间隔小于 τ_I,如式(7)所示。

$$DC_{t_1,t_2}^{t_3,t_4}=\langle LC_{t_1}^{t_2}\cdot L_{t_3}^{t_4}\mid t_3-t_2\leqslant\tau_I\rangle \quad (7)$$

(3)第 III 层的鼠标行为

第 III 层的鼠标行为是在鼠标移动的基础上完成的一些鼠标行为,包括鼠标的移动加左键单击行为(MM_LC)、移动加右键单击行为(MM_RC)、移动加左键拖拽行为(MM_LDD)、移动加右键拖拽行为(MM_RDD)以及移动加双击行为(MM_DC)。该层中的时间阈值定义如定义 5 所示。

定义 5 第 III 层中的时间阈值

①鼠标移动到左键单击的时间阈值(τ_{MLM}):鼠标移动事件和鼠标左键按下连接成一个行为的时间间隔。

②鼠标移动到右键单击的时间阈值(τ_{MRM}):鼠标移动事件和鼠标右键按下连接成一个行为的时间间隔。

在定义 5 的基础上,定义 6 给出第 III 层鼠标行为的描述。

定义 6 第 III 层中的鼠标行为定义

①鼠标移动加左键单击行为(MM_LC),由一系列的鼠标移动事件和一个在 τ_{MLM} 时间范围内发生的鼠标左键单击行为组成,如式(8)所示。

$$MM\_LC_{t_1}^{t_n}=\langle MMS_{t_1}^{t_{n-2}}\cdot LC_{t_{n-1}}^{t_n}\mid t_{n-1}-t_{n-2}\leqslant\tau_{MLM}\rangle \quad (8)$$

②鼠标移动加右键单击行为(MM_RC),由一系列的鼠标移动事件和一个在 τ_{MRM} 时间内发生的鼠标右键单击行为组成,如式(9)所示

$$MM\_RC_{t_1}^{t_n}=\langle MMS_{t_1}^{t_{n-2}}\cdot RC_{t_{n-1}}^{t_n}\mid t_{n-1}-t_{n-2}\leqslant\tau_{MRM}\rangle \quad (9)$$

③鼠标移动加左键拖拽行为(MM_LDD),由一系列的鼠标移动事件和一个鼠标左键拖拽行为组成,如式(10)所示。

$$MM\_LDD=\langle MMS_{t_1}^{t_{n-2}}\cdot LDD_{t_{n-1}}^{t_n}\mid t_{n-1}-t_{n-2}\leqslant\tau_{MLM}\rangle \quad (10)$$

④鼠标移动加右键拖拽行为(MM_RDD),由一系列的鼠标移动事件和一个鼠标右键拖拽行为组成,如式(11)所示。

$$MM\_RDD=\langle MMS_{t_1}^{t_{n-2}}\cdot RDD_{t_{n-1}}^{t_n}\mid t_{n-1}-t_{n-2}\leqslant\tau_{MRM}\rangle \quad (11)$$

⑤鼠标移动加双击行为(MM_DC),由一系列的鼠标移动事件和一个鼠标双击行为组成,如式(12)所示。

$$MM\_DC_{t_1}^{t_n}=\langle MMS_{t_1}^{t_{n-2}}\cdot DC_{t_{n-1}}^{t_n}\mid t_{n-1}-t_{n-2}\leqslant\tau_{MLM}\rangle \quad (12)$$

2.3 鼠标行为特征值及其计算方法

2.3.1 鼠标移动特征值

从 2.2 节关于鼠标行为的定义中可以看到,除了左键单击(LC)、右键单击(RC)、左键双击(DC),绝大多数的行为中都包括鼠标移动序列。因此本节将重点介绍关于鼠标移动的特征值。

对于一次鼠标移动序列(MMS),可以通过在捕获过程中获得的基础数据整理出 3 个向量来表示这个鼠标移动序列。这 3 个向量是横坐标向量、纵坐标向量以及捕获这些移动事件的时间向量,这些向量可公式化表示为:

①$x_i,i=1,\cdots,n$,在时间 t_i 处的横坐标

②$y_i,i=1,\cdots,n$,在时间 t_i 处的纵坐标

③$t_i,i=1,\cdots,n$,捕获鼠标移动事件的时间点

通过以上这些向量信息,可以计算出相关特征值,包括时间上和空间上的特征信息。

鼠标移动序列移动的距离即从开始到第 i 个点的路径长度 S_i 如式(13)所示。

$$S_i=\sum_{k=1}^{i-1}\sqrt{\delta_{x_k}^2+\delta_{y_k}^2},S_1=0 \quad (13)$$

这里,δ_{x_k} 与 δ_{y_k} 的表示如式(14)所示。

$$\delta_{x_k}=x_{k+1}-x_k,\delta_{y_k}=y_{k+1}-y_k \quad (14)$$

(1)时间上的特征值

时间上的特征值定义为 9 个向量,这 9 个向量由原始数据得到。本文对这 9 个向量给出下面的说明。

①x:由 $x_i\cdots x_n$ 值作为输入的向量

②y:由 $y_i\cdots y_n$ 值作为输入的向量

③t:由 $t_i\cdots t_n$ 值作为输入的向量

④v_x:水平方向速率,$v_x=\delta x/\delta t$

⑤v_y：竖直方向速率，$v_y=\delta y/\delta t$

⑥v：切向速率，$v=\sqrt{v_x^2+v_y^2}$

⑦$\dot{v}$：切向加速度，$\dot{v}=\delta v/\delta t$

⑧$\ddot{v}$：切向加加速度，$\ddot{v}=\delta\dot{v}/\delta t$

⑨w：角速率，$w=\delta\theta_i/\delta t$

在9个向量中，前3个向量是由原始数据直接得到，后6个向量由原始数据计算得到。

(2)空间上的特征值

空间上的特征值是由原始数据经过处理之后得到的关于鼠标移动序列的新的横纵坐标向量计算而来的。具体处理鼠标移动序列的方法是：对原始数据进行3次样条插值运算，从而得到平滑的曲线，然后对这个曲线进行等距离的抽样，最后得到新的向量(x_i', y_i')。

关于空间上的特征值，本文定义6个向量，这6个向量是由插值抽样后的数据得到的。对6个特征向量作出如下描述。

①x'：横坐标向量

②y'：纵坐标向量

③S'：从起点到结束的移动距离

④θ：移动角度

$$\theta_i=\arctan*(\frac{\delta y_1}{\delta x_1})+\sum_{j=1}^{i}\delta\theta_j$$

$$\delta\theta_j=\delta\arctan*(\frac{\delta y_i}{\delta x_j})$$

⑤c：曲率，$c=\delta\theta/\delta s$

⑥Δc：曲率变化率，$\Delta c=\delta c/\delta s$

(3)其他特征值

除了时间和空间两方面的特征值需要计算外，还有一些其他的特征值需要计算，包括移动时间t、移动距离S_n、直线度(straightness)、关键点(critical point)、路径抖动(jitter)、轨迹质量中心(trajectory center of mass)、散射系数(scattering coefficient)、势差(moment)，具体的计算公式如表1所列。对于向量$x', y', v_x, v_y, v, \dot{v}, \ddot{v}, w, \theta, c, \Delta c$，需要计算它们的最大值、最小值、平均值、标准差、区间长度。

表1 相关计算公式

特征名称	特征计算公式
移动时间	t_n-t_1
移动距离	$S_n;S_1=0$
直线度(S)	$\frac{\sqrt{(x_1-x_n)^2+(y_1-y_n)^2}}{s_n}$
关键点(CP)	$\sum_{i=1}^{n}Z_i, Z_i=\begin{cases}1, & \Delta c_i=0 \text{ and } \lvert c_i\rvert>\alpha\\ 0, & \text{otherwise}\end{cases}$, for $\alpha>\frac{\pi}{10}\frac{\text{rad}}{\text{pixel}^2}$
路径抖动(J)	S'/S_n
轨迹质量中心(TCM)	$TCM=\frac{1}{s_n}\sum_{t=1}^{n-1}t_{i+1}\sqrt{\delta x_i^2+\delta y_i^2}$
散射系数(SC)	$SC=\frac{1}{s_n}\sum_{t=1}^{n-1}t_{i+1}^2\sqrt{\delta x_i^2+\delta y_i^2}-TCM^2$
第三、第四势差 M_3、M_4	$M_k=\frac{1}{s_n}\sum_{t=1}^{n-1}t_{i+1}^k\sqrt{\delta x_i^2+\delta y_i^2}$ where k=3,4

2.3.2 鼠标特征值的处理方法

本文使用的特征值从取值范围角度来说属于连续型的变量，因此为了使用这些特征值训练分类模型，需要对这些特征值进行处理。

(1)距离离散化。在简单的双击或者单击行为中，一般情况下是没有鼠标位置变化的，因此本文在考虑这个特征值的时候只考虑鼠标位置是否发生了变化，而不考虑鼠标位置变化了多少。根据以下规则进行离散化处理，如果在单击或者双击行为中有位置的变化，则记为1；没有位置的变化，则记为2。这条离散化规则应用在TDC、FCD、ID、SCD、DC等特征值上。

(2)区间划分。其他的一些特征值，比如移动时间等，属于连续型的，不能直接应用在分类模型的训练上，因此需要将这些特征值进行一定的划分。将数据分成若干个区间，使用区间标号对落在同一个区间范围内数据进行标记，同一区间内的数据不再具有特征区分度。在本文中，对一个连续型的特征属性采用以最大信息增益为标准的划分方法，该方法包括两个部分，第一部分为计算待划分的区间，第二部分为求分割点集。

第一个部分根据每个用户在该特征属性上的高斯分布(μ,σ^2)计算出用户特征数据的大体分布区间。使用高斯分布来衡量用户数据的大概分布是出于两方面的考虑，从概率统计学角度来说，实际中大量的随机变量都服从或者近似服从高斯分布，同时又综合考虑了行为认证所具有的特点。本文中设定的数据范围为$(\mu-1.96*\sigma,\mu+1.96*\sigma)$，根据高斯分布，理论上该区间范围内数据占所有数据的95.449974%。最后计算所有合法用户区间的并集合，该区间为所求的待划分的区间。这种方法可以有效地减少数据内异常点对离散化数据的影响。

第二部分是在第一部分求得的区间的基础上进行以最大信息增益为标准的区间划分，最后得到区间内的分割点的集合。本文的划分标准是使得划分结束后区间的信息增益最大。采用的方法是对区间内所有的数据点进行升序排序，得到序列$(x_1, x_2, x_3, \cdots, x_n)$，在此序列内取每两个相邻点的中点作为待选的划分点，从中选取可以使得划分后区间信息增益最大的点作为此次划分的点，然后选择划分后两部分中熵较大的那部分继续划分，直到满足设定的停止划分条件。本文设定的停止条件为达到设置的区间个数(该区间个数不能大于现有的合法用户个数)或者区间不可继续划分。

2.4 基于随机森林的分类器构造方法

本节利用基于随机森林的方法来构建分类器，进而达到对训练阶段收集到的用户鼠标行为进行分类判断的目的。

在本方案中，每一种鼠标行为对应一个随机森林分类器，每一种类型的鼠标行为数据通过矩阵的形式表现出来，矩阵的每一列对应于这个鼠标行为的一个特征类型，矩阵的每一行对应一个用户该类型行为的实例即特征值向量，每一行数据通过用户ID进行标识。

对于一个$N*M$的行为矩阵，N代表矩阵的行向量个数，表示为该行为类型的实例个数，这些实例来源于该系统的合法用户；M表示列向量的个数，表示这种行为类型所具有的特征值个数。一个鼠标双击行为矩阵如图2所示。

$$\begin{pmatrix}2 & 168\text{ms} & 2 & 104\text{ms} & 2 & 128\text{ms} & 2\\ 2 & 135\text{ms} & 2 & 120\text{ms} & 2 & 136\text{ms} & 2\\ 1 & 127\text{ms} & 1 & 319\text{ms} & 1 & 120\text{ms} & 2\\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots\\ 3 & 96\text{ms} & 2 & 88\text{ms} & 2 & 185\text{ms} & 2\end{pmatrix}$$

图2 鼠标双击行为矩阵

矩阵第一列表示的是行向量标识，为合法用户的ID值，其余列表示对应该类型的鼠标行为特征值，在鼠标双击行为矩阵里，M为6，特征值名字依次为FCT、FCD、IT、ID、SCT、SCD。

每个行为类型对应一个随机森林分类器，随机森林内的每一棵决策树都是通过该行为矩阵里的数据进行训练得到的。每一棵决策树的构建都有两个过程：采样和分裂。

采样过程包括两种类型的采样过程，一种是对行为矩阵的行向量进行有放回的抽样，也就是说采样得到的样本集合里面可能有重复的样本。这样使得在训练时，每一棵树的输入样本都不是全部样本，进而减少了过度拟合情况的出现。第二种采样是对行为矩阵的列向量进行采样，从 M 个特征值中选择 m（$m \ll M$）个特征值，这里 m 一般选择 M 的平方根。对采样之后的数据使用完全分裂的方式建立决策树。

在决策树分裂过程中，每一个分裂节点处属性的选择采用信息增益比作为分裂的依据。选择使信息增益比最大的鼠标行为特征值作为该节点分裂的属性。首先给出信息增益的概念。

若属性 α 按照规则将样本集 T 划分为 $T_1, T_2, \cdots, T_m$，共 m 个子集，那么信息增益 $Gain(\alpha)$ 的计算公式如式(15)所示。

$$Gain(\alpha)=Entrop(T)-\sum_{i=1}^{m}\frac{|T_i|}{|T|}\times Entrop(T_i) \tag{15}$$

$|T|$ 为 T 的样本个数；$|T_i|$ 为子集 T_i 的样本个数；$Entrop(T)$ 表示样本集合的信息熵，其定义如式(16)所示。

$$Entrop(T_u)=-\sum_{i=1}^{s}P(u_i)\times\log_2(P(u_i)) \tag{16}$$

T_u 表示一个样本集合，共包含 s 个类别，u_i 表示其中的一个类别的样本集合，$P(u_i)$ 表示这个类别出现的频率，定义如式(17)所示。

$$P(u_i)=\frac{|u_i|}{|T_u|} \tag{17}$$

信息增益表示的是整体样本集合 T 的熵相对于用 α 分类 T 后得到 m 子类后各个子类的加权熵和的变化量。权值就是子集 T_i 样本在全集中所占的比例。

信息增益比的定义如式(18)所示。

$$GainRation(\alpha)=\frac{Gain(\alpha)}{SplitInfo(\alpha)} \tag{18}$$

其中，$Gain(\alpha)$ 是之前定义的信息增益；$SplitInfo(\alpha)$ 表示为拆分信息，即把 T 分成 m 个部分而生成的潜在信息。$SplitInfo(\alpha)$ 的计算公式如式(19)所示。

$$SplitInfo(\alpha)=-\sum_{i=1}^{m}\frac{|T_i|}{|T|}\times\log_2\frac{|T_i|}{|T|} \tag{19}$$

在每个节点，根据信息增益比做出分裂属性选择，选取能够获得信息增益比最大的特征值作为分裂属性，对样本空间进行分裂。决策树的生长停止条件有两条：第一节点处的样本集合都属于同一类别；第二节点处没有可以选择的属性值进行下一步的分裂。

对于分裂停止的两种情况，本文选用多数原则作为标识叶子节点类别的准则，即在当前叶子节点内的数据集 D_j 包括的所有类别中，占有最大比例的类别就是该叶子节点的类别标识。

随机森林内部的决策树构造完成后，不需要进行剪枝操作，因为在构造阶段的两次抽样已经在一定程度上避免了数据与分类模型的过度拟合。按照上述方法构造出一种类型的鼠标行为分类器中的全部决策树，至此随机森林分类器构造结束。同样可构造出其他鼠标类型的分类器，将构造好的分类器存入模式数据库里，供决策阶段使用。

2.5 基于分类决策模型的用户身份认证

本文提出的分类决策模型包括两层，第一层是分类器判断层，第二层是决策层。基于分类决策模型的用户身份认证过程如图 3 所示。

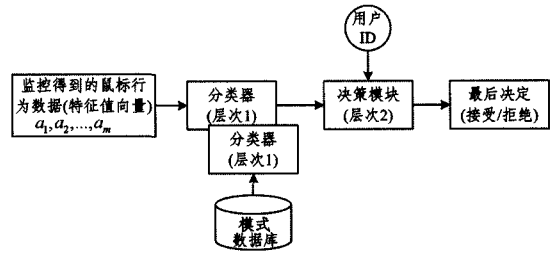

图 3 基于分类决策模型的用户身份认证过程

图 3 中的用户 ID 是通过用户名和密码与服务器建立合法连接的用户 ID 值。$a_1, a_2, \cdots, a_m$ 在这里是指从建立连接处监控得到的鼠标行为数据，也就是代表每一个鼠标行为的特征向量。

为进一步介绍分类决策模型，首先给出如下符号描述。

假设系统内有 n 个合法用户，$U=\{u_1, u_2, \cdots, u_n\}$ 代表经过训练的合法用户集合，$A=\{a_1, a_2, \cdots, a_m\}$ 代表一组监控到的鼠标行为数据序列。

2.5.1 分类层

在分类决策阶段，从模式数据库里取出 $A=\{a_1, a_2, \cdots, a_m\}$ 用于行为分类。服务器端在收到 $A=\{a_1, a_2, \cdots, a_m\}$ 之后，将其送入对应的行为类型分类器内。对于一个行为 a_j，其对应的分类器预测每个经过训练的用户 u_i 完成行为 a_j 的可能性，记为 $\hat{P}(u_i|a_j)$，这个值通过随机森林计算得出。

本文采用的方法是均权投票的方法，即随机森林内每一棵决策树投出的票具有相同的权重。设一个行为的随机森林分类器内共有 W 棵决策树，对于行为 a_j，有 W_i 棵树的分类结果是用户 u_i，则 $\hat{P}(u_i|a_j)$ 如式(20)所示。

$$\hat{P}(u_i|a_j)=\frac{W_i}{W} \tag{20}$$

定义 $P^{apr}(u_i|a_j)$ 为先验概率，这个概率是从训练数据集中得到的。T_j 是标记来源于合法用户的行为类型 a_j 的训练集合，同时 $T_{ij}=\{t_{ij}^1, t_{ij}^2, \cdots, t_{ij}^{m_{ij}}\}$ 表示用户 u_i 完成行为类型 a_j 的 m_{ij} 个实例，则 $P^{apr}(u_i|a_j)$ 如式(21)所示。

$$P^{apr}(u_i|a_j)=\frac{m_{ij}}{|T_j|} \tag{21}$$

由于在行为收集阶段，每一种类型的行为数据的个数有很大的不同，即 m_{ij} 的值有很大的不同，因此由此数据集建立起来的分类器可能会给出一个带偏向的计算结果。为了解决这个问题，将对结果进行正态化的操作，正态化后的概率为 $P^{norm}(u_i|a_j)$，其计算公式如式(22)所示。

$$P^{norm}(u_i|a_j)=\frac{\frac{\hat{P}(u_i|a_j)}{P^{apr}(u_i|a_j)}}{\sum_{t=1}^{n}\frac{\hat{P}(u_t|a_j)}{P^{apr}(u_t|a_j)}} \tag{22}$$

最后给出用于第二层计算不带偏向的用户 u_i 完成行为 a_j 的概率 $P^{post}(u_i|a_j)$，其计算公式如式(23)所示。

$$P^{post}(u_i|a_j)=\frac{P^{norm}(u_i|a_j)}{\sum_{t=1}^{n}P^{norm}(u_t|a_j)} \tag{23}$$

2.5.2 决策层

在决策层里，将结合第一层给出的分类运算结果，输出对于一组监控到的行为序列 $A=\{a_1, a_2, \cdots, a_m\}$ 的最后判断结果。

首先，决策层结合分类层给出的每个行为的分类计算结果 $P^{post}(u_i|a_j)$，计算得到每个用户完成该行为序列的可能性 $P^{post}(u_i|a_1,a_2,\cdots,a_m)$，行为序列中的每一个行为都是相对独立的，如式(24)所示。

$$P^{post}(u_i|a_1,a_2,\cdots,a_m)=\frac{\sum_{j=1}^{m}P^{post}(u_i|a_j)}{\sum_{i=1}^{n}\sum_{j=1}^{m}P^{post}(u_i|a_j)} \tag{24}$$

如果 $P^{post}(u_i|a_1,a_2,\cdots,a_m)$ 高于一个设定好的阈值，则认为该行为序列来源于用户 u_i，其计算公式如式(25)所示。

$$Final\ Decision(\{a_1,a_2,\cdots,a_m\}\in u_i)=\begin{cases}\text{Yes}, & P^{post}(u_i|a_1,a_2,\cdots,a_m)\geqslant\lambda\\ \text{No}, & \text{otherwise}\end{cases} \tag{25}$$

3 实验分析

3.1 数据采集与预处理

用户鼠标行为数据的采集是在用户的日常工作中进行的，每个用户都在各自的计算机上安装一个可以被动监控并记录用户鼠标行为的客户端软件，并将采集的数据自动送到采集服务器。本文共采集了 20 个计算机用户在 2 个月内的鼠标行为数据。参与者的计算机的显示器均为 19 英寸 LCD，显示分辨率都为 1366 * 768，内存均为 2GB，其他的硬件配置略有不同。软件系统使用的是 Windows 不同版本的操作系统：Windows XP(9 台)和 Win 7(11 台)。采集到的输入数据包括一系列的鼠标动作、屏幕坐标、系统时间、进程信息等。

在本次实验中进行实验的鼠标特征值数据从行为数据库中随机抽样获得，抽样不同行为数据数量的比例依赖于数据库内已经收集到的不同行为类型的数量比例，以便模拟真实监控到的数据比例。在本文的实验中，不同的用户选择不同的系统阈值来进行身份认证，系统阈值的设定由实验获得。

3.2 实验分析

对于一个基于用户鼠标行为的身份认证方法来说，最重要的性能参数就是错误拒绝率(FRR(False Rejection Rate)：是指把合法用户当作非法用户，并拒绝其访问计算机的信息资源的概率)和错误接受率(FAR(False Acceptance Rate)：指非法用户被当成合法用户，并授权非法用户访问计算机系统资源的概率)。本文通过这 2 个性能参数对所提出的认证方法进行分析。

(1)FRR 和 FAR 与用户行为数量的关系

图 4、图 5 分别表示 FRR、FAR 随行为数量的变化关系。从图 4 可以看出，随着抽样数据的增长，FRR 呈现下降的趋势，在初始阶段，FRR 下降趋势明显，当数据达到一定量时，FRR 值会稳定在一个范围内。从图 5 可以看出，随着抽样数据的增长，系统的 FAR 同样呈现下降趋势。实验结果表明，对于一个用户来说，在用户身份认证阶段，一次认证过程中监控到的用户鼠标行为数据量会影响系统对身份的认证结果。

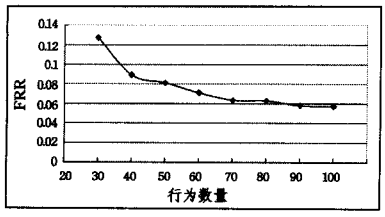

图 4 FRR 随行为数量的变化

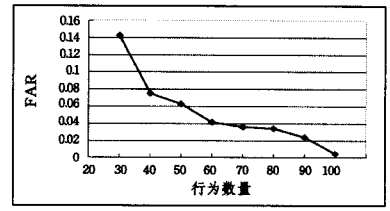

图 5 FAR 随行为数量的变化

(2)FRR 和 FAR 之间的关系

在本次实验中选取行为数据量为 70 作为测试标准，对于合法用户，采用不同的阈值设定，实验结果如图 6 所示。从实验结果可以看到，FRR 与 FAR 呈反比的关系，因此对于具有不同安全要求的系统来说，可以通过调整安全的阈值，使其更加符合系统要求，比如对于一个安全要求比较严格的系统来说，可以调整为较大的阈值，此时系统的 FAR 较低，表明错误接受的概率较小。

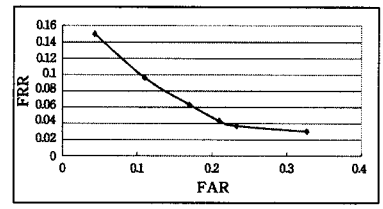

图 6 FRR 与 FAR 之间的变化关系

(3)内部用户和外部用户进行身份欺骗时的 FRR 和 FAR

内部用户是指已经加入到系统内并通过模型训练的合法用户；外部用户是指未加入到系统内，即系统分类模型中不包含该用户的特征信息。在本次实验中两类用户都对合法用户进行冒充，行为数量基准为 70，两种类型的数据均来自实验参与者所安装的客户端收集软件，实验结果如图 7 所示。

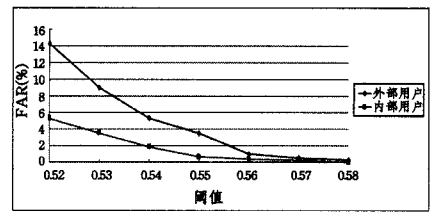

图 7 系统性能参数对比

从实验结果可以看出，在对同一个用户进行身份冒充时，在阈值可以接受的合理设定前提下，阈值设定比较严格时，两者的 FAR 比较接近，阈值设定较低，两种类型的用户在 FAR 上存在差别。最后，对加入到系统中的合法用户进行交叉的认证实验工作，行为数量基准设定为 70，实验结果表明得到的平均 FRR 为 11.63%，FAR 为 3.96%。

结束语 本文提出了一种基于用户鼠标行为的身份认证的方法。首先，给出了鼠标事件及行为的相关定义，并给出了对应每种鼠标行为类型的特征值及其计算方法。为避免数据的过度拟合和数据噪声的影响，采用随机森林分类器作为鼠标行为分类的工具。在身份认证阶段，结合用户鼠标行为，采用层次结构的分类决策模型对用户身份进行认证。最后，对本文提出的方法进行实验分析，结果表明本方法具有较好的错误拒绝率和错误接受率。

参 考 文 献

[1] Wang Yi, Hu Jian-kun. Global Ridge Orientation Modeling for

Partial Fingerprint Identification [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2011, 33(1): 72-87

[2] Jing Xiao-yuan, Li Sheng, Zhang D, et al. Optimal subset-division based discrimination and its kernelization for face and palmprint recognition[J]. Pattern Recognition, 2012, 45(10): 3590-3602

[3] Su Yu, Shan Shi-guang, Chen Xi-lin, et al. [J]. Journal of Software, 2010, 21(8): 1849-1862(in Chinese)
苏煜,山世光,陈熙霖,等. 基于全局和局部特征集成的人脸识别[J]. 软件学报, 2010, 21(8): 1849-1862

[4] Nigam A, Gupta P. Iris recognition using consistent corner optical flow [C]//Proc of 11th Asian Conference on Computer Vision. 2012: 358-369

[5] Lee P J-W, Choi S-S, Moon P B-R. An evolutionary keystroke authentication based on ellipsoidal hypothesis space [C]// Proc of the 9th Annual Conference on Genetic and Evolutionary Computation. 2007: 2090-2097

[6] Ahmed, TraoreIssa A A. Biometric recognition based on free-text keystroke dynamics [J]. IEEE Transactions on Cybernetics, 2014, 44(4): 458-472

[7] Ahmed A A E, Traore I. Detecting computer intrusions using behavioral biometrics[C]// Proc of 3rd Annual Conference on Privacy. Security, 2005: 91-98

[8] Jorgensen Z, Yu T. On mouse dynamics as a behavioral biometric for authentication[C]//Proc of the 6th ACM Symposium on Information, Computer and Communications Security. 2011: 476-482

[9] Gamboa H, Fred A L N, Jain A K. Webbiometrics: User verification via web interaction[C]// Proc of 2007 Biometrics Symposium. 2007: 1-6

[10] Pusara M, Brodley C E. User re-authentication via mouse movements[C]// Proc of the 2004 ACM workshop on Visualization and data mining for computer security. 2004: 1-8

[11] Bours P, Fullu C J. A login system using mouse dynamics[C]// Proc of 5th International Conference on Intelligent Information Hiding and Multimedia Signal Processing. 2009: 1072-1077

[12] Schulz D A. Mouse curve biometrics [C]// Proc of 2006 Biometrics Symposium. 2006: 79-83

[13] 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

[14] Shen Chao, Cai Zhong-min, Guan Xiao-hong, et al. User authentication and monitoring based on mouse behavioral features[J]. Journal on Communications, 2010, 31(7): 68-75(in Chinese)
沈超,蔡忠闽,管晓宏,等. 基于鼠标行为特征的用户身份认证与监控[J]. 通信学报, 2010, 31(7): 68-75

(上接第 123 页)

[26] Jeh G, Widom J. SimRank: A measure of structural context similarity[C]//Proceedings of the ACM SIGKDD 2002. New York: ACM Press, 2002: 538-543

[27] Zhou T, Lv L, Zhang Y-C. Predicting missing links via local information[J]. European Physical Journal B, 2009, 71(4): 623-630

[28] Lv L, Jin C-H, Zhou T. Similarity index based on local paths for link prediction of complex networks[J]. Physical Review E-Statistical, Nonlinear, and Soft Matter Physics, 2009, 80(4): 046122

[29] Liu W-P, Lv L. Link Prediction Based on Local Random Walk [J]. European Physics Letter. , 2010, 89(5): 58007

[30] Rao Jun, Wu Bin, Dong Yu-xiao. Parallel Link Prediction in Complex Network Using MapReduce[J]. Journal of Software, 2012, 23(12): 3175-3186(in Chinese)
饶君,吴斌,东昱晓. MapReduce 环境下的并行复杂网络链路预测[J]. 软件学报, 2012, 23(12): 3175-3186

[31] Dong Yu-xiao, Ke Qing, Wu Bin. Link Prediction Based on Node Similarity[J]. Computer Science, 2011, 38(7): 162-164(in Chinese)
东昱晓,柯庆,吴斌. 基于节点相似性的链接预测[J]. 计算机科学, 2011, 38(7): 162-164

[32] http://www.linkprediction.org/index.php/link/resource/data

[33] Latora V, Marchiori M. Efficient behavior of small-world networks[J]. Phys. Rev. Lett. , 2001, 67: 198701-198704

[34] Watts D J, Strogatz S. Collective dynamics of 'small-world' networks[J]. Nature. , 1998, 393(6684): 440-442

[35] Newman M E J. Assortative mixing in networks[J]. Phys. Rev. Lett. , 2002, 89(20): 208701-208705

[36] Newman M E J. Scientific collaboration networks. I. network construction and fundamental results [J]. Physical Review E, 2001, 64: 0161311-061317

[37] Newman M E J. Scientific collaboration networks. II. Shortest paths, weighted networks, and centrality [J]. Physical Review E, 2001, 64: 0161321-0161327

[38] Liu Ai-fen, Fu Chun-hua, Zhang Zeng-ping, et al. An Empirical Statistical Investigation on Chinese Mainland Movie Network [J]. Complex Systems and Complexity Science, 2007, 4(3): 10-16

[39] Robins G, Alexander M. Small worlds among interlocking directors: network structure and distance in bipartite graphs [J]. Computational & Mathematical organization Theory, 2004, 10(1): 69-94

[40] Battiston S, Catanzaro M. Statistical properties of corporate board and director networks [J]. European Physics Journal B, 2004, 38(2): 345-352

[41] Chen Wen-qin, Lu Jun-an, Liang Jia. Research in Disease-Gene Network Based on Bipartite Network Projection[J]. Complex Systems & Complexity Science, 2009, 6(1): 13-19

[42] Ergun G. Human sexual contact network as a bipartite graph [J]. Physica A, 2002, 308(1-4): 483-488

[43] Lambiotte R, Ausloos M. Uncovering collective listening habits and music genres in bipartite networks [J]. Physical Review E, 2005, 72(6): 066107

[44] Le Blond S, Guillaume J L, Latapy M. Clustering in P2P exchanges and consequences on performances[C]// Castro M, Renesse R, eds. Peer-to-Peer Systems IV. Berlin: Heidelberg, 2005, 193-204