基于键盘行为进行用户识别的方法与应用

王昱杰 赵培海 王咪咪
(同济大学电子与信息工程学院 上海 201800)

摘 要 基于键盘行为的已有研究大多限制在实验环境下,并限定用户的输入为特定字符串,或只在有限实验参与者中进行识别,从而导致其无法应用在真实的环境中。为了克服已有研究的局限,解决真实环境中正负样本不均衡和负样本缺失的问题,对真实环境下一千万条用户击键行为进行分析,提出了一种基于键盘行为进行用户识别的方法。该方法使用击键持续时间和击键间隔时间描述用户的击键行为,通过实验选择马氏距离进行用户击键行为相似度的对比,并对未知击键行为进行预测。实验结果表明,该方法可获得80%的预测准确率。

关键词 用户识别,身份校验,击键行为,马氏距离,向量相似性
中图法分类号 TP181 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2015.11.042

Method of User Identification Based on Keystroke Behavior and its Application

WANG Yu-jie ZHAO Pei-hai WANG Mi-mi
(School of Electronics and Information Engineering, Tongji University, Shanghai 201800, China)

Abstract Most researches on keyboard behavior have been limited in experimental environment, which force user to input a specific string or limite the number of participants, so the result of research can't apply into real production environment. To overcome this limitation and resolve the problem of the unbalanced samples between positive and negative and the problem of lacking negative samples, ten million keystroke behaviors were analyzed, and a method for user identification was introduced. The duration and the interval of keystroke describing the user's behavior are used, and the similarity of different keystrokes is measured by Mahalanobis distance. The result of experiment shows that our method gets an accuracy of 80%.

Keywords User identification, Identify verification, Keystroke behavior, Mahalanobis distance, Similarity of vector

1 引言

互联网的发展促使对互联网应用的安全性要求也随之提高。传统的密码等安全机制由于面临着破解、泄露的风险,已无法满足当下日益增长的安全需求。另一方面,基于身份特征与行为特征的用户信息在安全要求方面获得了显著效果[1-6]。用户信息包括两种,第一种是以虹膜、指纹、掌纹以及声纹等为代表的生物信息。这一类生物信息具有很强的个人标识作用,如利用指纹的高精准度身份识别进行手机解锁、身份认证、基于图像识别所进行的人脸识别等。但此类生物信息需要额外的硬件设备进行采集,指纹需要通过指纹传感器进行采集,人脸识别需要摄像头进行图像采集。对采集设备的依赖性限制了此类安全措施的应用。另外一种用户特征信息是用户的行为模式特征,包括用户的键盘行为模式、鼠标行为模式、用户在浏览网页时的偏好等。这些行为模式在不同的用户群体之间具有明显的差异,并且无需额外的硬件设备进行采集。其中,键盘行为由于其简单的采集成本和较好的应用性得到了广泛的研究。

键盘行为最早的研究是基于键盘行为的静态认证,即用户在登录时输入特定的字符串,分析该字符串的特征以进行安全认证[10]。静态认证在有限的输入集中获得了较高的识别准确率,但该认证具有一定局限性,当用户输入预设内容之外的字符串时,识别率较低。为了克服静态认证的局限性,研究者们对静态认证进行了扩展,即允许用户输入任意的字符串,包括基于键盘行为的动态认证[11-13]。在现有的键盘行为研究中,特征信息从击键的按下和弹起等时间特性上进行提取。最常见的特征包括单个按键按下的持续时间、相邻按键的时间间隔。文献[11]通过欧氏距离和朴素贝叶斯分类器,对连续的击键行为进行用户无感知的实时认证。通过基于欧氏距离的计算,利用更少的输入内容进行认证,其所使用的数据来自55名志愿者在一年内使用电子邮件时的输入行为信息,在大数据量和更多用户的情况下,实验的效果未知。文献[8]将马氏距离应用在击键行为中,使得模型能够更好地适应击键行为数据规模的变化,处理数据中的异常值,降低特征之间的相互影响和冗余。其数据集采用了卡耐基梅隆大学所提供的键盘动态特征标准数据集,该数据集收集了51名用户在重复400次密码输入时的击键特征。文献[17]将击键时的动态特征与密码结合,通过神经网络的训练提高Web应用的安全性。其数据来源于25名用户的密码输入,且限定密码长度为7个字符,重复输入密码150~400次,使用最后75次进行

到稿日期:2014-11-17 返修日期:2015-03-30 本文受港澳台科技合作专项(2013DFM10100)资助。
王昱杰(1989-),男,硕士生,主要研究方向为用户行为识别、用户身份认证等,E-mail:wangfan@foxmail.com;赵培海(1989-),男,博士生,主要研究方向为用户行为分析、电子交易安全和数据挖掘等;王咪咪(1985-),女,博士生,主要研究方向为可信服务、一致性检测、Petri网理论等。

实验。文献[18]使用了曼哈顿距离进行键盘行为的度量，通过数据流的方式描述击键之间的时间延迟，进行用户认证。其实验在33名用户中进行，随机选择其中6名用户作为目标对象，剩余27名用户输入与6名用户相同的用户名、密码，进行仿造。针对用户击键时输入模式可能的变化，文献[19]提出了一种持续再训练的方法，以容纳用户输入模式的变动，即使用固定大小的滑动窗口选择最近几次的击键行为进行训练，或使用可变大小的滑动窗口选择历史上所有的击键行为进行训练，在训练时使用了K-means算法。其算法应用在一个由21名用户的击键行为组成的数据集上，每个用户具有的击键行为从76到338条不等。文献[20]研究了最近邻分类器、高斯模型以及One-Class支持向量机在击键行为数据集上的作用。其数据来自21名用户输入密码时的键盘行为，密码长度为6到10个字符不等，每个用户重复输入150～400次，选择最后的75次进行实验。这些研究通过相似性度量方法的改进和机器学习方法的改进，使用键盘行为进行用户识别在特定的实验环境下取得了较好的成果。

但同时这些实验的数据集都限定在少数用户，且在实验环境下进行采集。然而在真实的应用环境中，用户输入的内容、用户所使用的键盘与操作系统等因素都无法控制，导致针对某个用户的识别可能产生正负样本不均衡和负样本缺失的问题。

针对以上情况，使用国内某互联网公司在真实环境中的数据，对一百多万名用户的一千万条键盘行为数据进行分析，在已有研究的基础上，提出了一种运用键盘行为进行用户识别的方法。在键盘行为中，由于击键行为的稳定性，同一用户不同击键行为之间的相似度比不同用户击键行为的相似度更高。基于这样的事实，从某一用户的已知击键行为出发，从击键行为中提取击键时间上的特征，并通过向量描述其击键行为。通过计算不同向量之间的距离来衡量不同击键行为之间的相似性。由向量之间的距离选择某一用户的击键行为阈值，当进行预测时，计算未知击键与已知击键行为之间的相似度。当相似度比阈值小时，接收该用户，预测其为正常用户；当相似度大于阈值时，拒绝该用户，预测其为非法用户。

本文第2节介绍了一些相关概念；第3节介绍了基于距离进行键盘行为识别的方法；第4节是键盘行为的实验及结论；最后总结全文。

2 相关概念

文献[8]对键盘行为中的基本特征进行了定义，包括击键被按下的时间(Press)、击键弹起的时间(Release)、单个击键的持续时间(Dwell Time)、相邻击键的间隔时间(Flight Time)。

借鉴已有的定义，本文给出定义1－定义9。

定义1(用户的击键序列) 设按键被按下的时刻为 P，按键弹起的时刻为 R。用户的一次输入序列中，按下和弹起时刻为：$P_1,R_1,P_2,R_2,\cdots,P_n,R_n$，则用户的击键序列为 $v_0=\{P_1,R_1,P_2,R_2,\cdots,P_n,R_n\}$。击键序列表示用户在击键时每个击键事件所发生的时间先后顺序。

定义2(用户击键持续时间序列) 设每次击键时按键被按下的时刻为 P，按键弹起的时刻为 R。用户此次击键的间隔时间为 $R-P$。在用户的一次输入序列中，击键持续时间序列为：$v_1=\{PR_1,PR_2,\cdots,PR_n\}$，其中 $PR_i=R_i-P_i(i=1,\cdots,n)$。

定义3(用户击键间隔时间序列) 设相邻两次击键中，第一个按键弹起的时间为 R，第二个击键按下的时间为 P，则此相邻击键的击键间隔时间为 $P-R$。在用户的一次输入序列中，击键间隔时间序列为 $v_2=\{RP_1,RP_2,\cdots,RP_n\}$，其中 $RP_i=P_{i+1}-R_i(i=1,\cdots,n-1)$。

定义4(用户击键行为的马氏距离) 设某次击键行为为向量 v_a，另一次击键行为为向量 v_b(该击键行为与第一次击键行为可能来自同一用户，也可能来自不同用户)。两次击键行为的马氏距离为：$D(v_a,v_b)=\sqrt{(v_a-v_b)^{\mathrm{T}}S^{-1}(v_a-v_b)}$。其中 v_a,v_b 均是 n 维向量，S 为 v_a 和 v_b 的协方差矩阵。$S=E[(v_a-\mu_a)(v_b-\mu_b)]$。其中 μ_a,μ_b 分别为 v_a,v_b 的均值。

定义5(用户击键行为的欧氏距离) 设某次击键行为为向量 v_a，另一次击键行为为向量 v_b。两次击键行为的欧氏距离为：$D(v_a,v_b)=\sqrt{(v_a-v_b)(v_a-v_b)^{\mathrm{T}}}$。

定义6(用户击键行为的曼哈顿距离) 设某次击键行为为向量 v_a，另一次击键行为为向量 v_b。两次击键行为的曼哈顿距离为：$D(v_a,v_b)=\sum_{k=1}^{n}|v_{ak}-v_{bk}|$。其中 v_{ak},v_{bk} 分别为向量 v_a,v_b 中的第 k 个元素。

定义7(击键行为相似度阈值) 设用户A的多次击键行为为集合 $V_A=\{v_1,v_2,\cdots,v_n\}$，可计算得到A的多次击键之间的相似度，设为集合 $s=\{D_1,D_2,\cdots,D_m\}$。在集合 s 中存在某个相似度 D_x，使得在集合中小于 D_x 的元素占集合所有元素的比例为分位数 k(k 为0到1之间的小数)，则击键行为相似度阈值为 D_x。

定义8(震荡幅度) 设已知用户A的击键行为为集合 $V_A=\{v_1,v_2,\cdots,v_n\}$，由 V_A 得到用户A的击键行为相似度阈值为 D_A。对未知击键行为 v_x 可分别计算其与用户A的已知击键行为的相似度，得到集合 $s_x=\{D_{1x},D_{2x},\cdots,D_{nx}\}$。

存在 $t(0<t<1)$，使得在集合中 s_x 小于阈值 D_A 的元素占集合所有元素的比例为 t，则震荡幅度为 t。

定义9(预测准确率) 设用户A具有两个击键行为的集合 V_{A1} 与 V_{A2}，将 V_{A1} 作为训练集计算得到用户A的击键行为相似度阈值，V_{A2} 作为测试集进行验证。V_{A2} 中的击键行为总数为 n，通过震荡幅度和阈值的调控，预测为用户A击键的数量为 m，则准确率为 m/n。即：预测准确率＝正确预测的击键行为数量/实际击键行为数量。

3 基于距离的键盘行为识别方法

已有键盘行为的研究中，无论静态认证或动态认证，针对击键行为的模型训练都需要足够的正、负样本，从正、负样本中学习获得足够的知识，从而对未知的击键行为进行预测。但在真实场景中，由于无法限定用户输入相同的内容，因此对于绝大部分的击键行为，没有与正样本对应的负样本数据，只有很少的击键行为具有负样本数据。例如，某一张信用卡被盗用时，输入内容均为信用卡号，以正常用户的击键行为作为正样本，盗用者的击键行为作为负样本。

3.1 可行性分析

由于负样本的缺失，导致无法使用分类器进行键盘行为的识别。为了解决样本缺失的问题，可以使用其他用户的键盘行为数据模拟某个用户的负样本。但使用模拟样本进行分类器的训练时，会使得分类器只能拟合模拟数据中的键盘行为模式，对于模拟样本之外的键盘行为，分类器无法准确识别，这使得通过模拟样本训练的分类器具有较大的局限性。

模拟样本无法取得良好的分类效果时，如果使用模拟样本，不同自然人之间的键盘行为是否存在明显的特征差异呢？我们从键盘行为数据中筛选了典型的案例，从 2 个维度区分了击键特征。

以单个按键的持续时间 $v_1=\{PR_1,PR_2,\cdots,PR_n\}$ 进行区分，如图 1 所示。

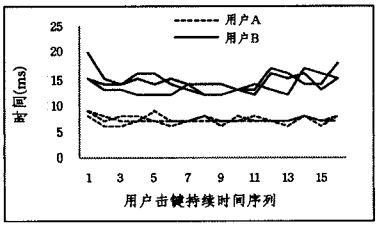

图 1　以击键持续时间进行用户行为的区分

图 1 中横坐标表示第 i 次击键，纵坐标表示该次击键的持续时间。从击键持续时间上可以看出，不同用户的击键持续时间差别较明显。用户 A 的多次输入中，击键持续时间稳定在 7～9ms 之间；而用户 B 的击键持续时间在 13～20ms 之间波动。这表明，用户 A 的每次击键时间比用户 B 短。

以相邻按键的间隔时间 $v_2=\{RP_1,RP_2,\cdots,RP_n\}$ 进行区分，如图 2 所示。

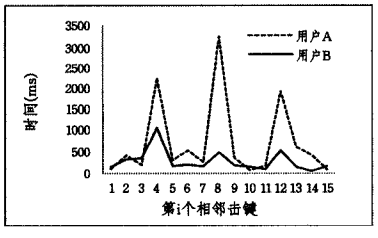

图 2　以相邻击键的间隔时间进行用户行为的区分

图 2 中横坐标表示第 i 个相邻击键，纵坐标表示该相邻击键的间隔时间。击键间隔可以反映出用户在输入时的数字组合方式，如图 2 中波峰出现的地方表示用户在输入时出现了较长时间的停顿和延迟。不同用户的折线图中，波峰的个数和出现的位置并不相同，表明不同用户在输入时的组合方式并不相同。图 2 中用户 A 在输入时出现了 3 个明显的波峰，表明用户 A 在输入时出现了较大的停顿，而用户 B 的输入更为均匀。

3.2　基于距离进行键盘行为识别的方法

正负样本的不均衡问题导致缺失负样本的信息，进而无法使用分类器进行击键行为的识别。通过 3.1 节中的可行性分析，可以看出不同的用户在击键行为上存在着显著的特征差异。由于击键行为的稳定性，同一用户的多次击键行为之间的相似度比不同用户的击键行为之间的相似度更高。基于此提出了如下方法。

针对某个已知用户的击键行为，计算得到该用户击键行为的阈值。当对未知的击键行为进行预测时，计算未知击键行为与所有已知击键行为的相似度，得到未知击键行为与已知击键行为距离的集合。当集合中小于阈值的元素比例大于震荡幅度时，认为新的击键行为的特征符合已知击键行为的特征，预测为合法用户；当集合中小于阈值的元素比例小于震荡幅度时，认为新的击键行为的特征不符合已知击键行为的特征，预测为非法用户。

基于以上思想，使用键盘行为进行用户识别的步骤如算法 1 所示。

算法 1　基于距离的键盘行为识别

输入：用户 A 的 n 次击键行为序列($v_1,v_2,\cdots,v_n$)与未知用户的某次击键行为 v_x。

输出：true 或 false。true 表示判定未知用户为 A 用户，false 表示判定未知用户不是 A 用户。

步骤：

1. 计算用户 A 击键序列中每次击键之间的两两距离，并将距离按照升序排列，得到距离($d_1,d_2,\cdots,d_m$)，其中 $m=\frac{n*(n-1)}{2}$。
2. 从距离序列($d_1,d_2,\cdots,d_m$)中选择阈值 d_t。
3. 计算未知用户击键行为 v_x 与用户 A 击键行为序列($v_1,v_2,\cdots,v_n$)的距离，得到($d_{1x},d_{2x},\cdots,d_{nx}$)共 n 个距离。
4. 计算($d_{1x},d_{2x},\cdots,d_{nx}$)中小于阈值 d_t 的百分比 t_0。
5. 震荡幅度为 t，比较 t_0 与 t 的大小。
6. 如果 t_0 大于或等于震荡幅度 t，输出 true，判定未知用户为 A 用户；如果 t_0 小于震荡幅度 t，输出 false，判定未知用户不是 A 用户。

阈值的计算方法如下：用户 A 的所有击键行为为集合 $V_A=\{v_1,v_2,\cdots,v_n\}$，共对应 n 次不同的击键序列。计算集合 V_A 内所有向量两两之间的距离，在实验室进行了 3 组实验，第一组计算向量之间的马氏距离，第二组计算向量之间的欧氏距离，第三组计算向量之间的曼哈顿距离。设 P_{ij} 为 v_i 和 v_j 之间的距离，集合 V_A 内有不同向量之间的距离 $D=\{D_{12},D_{13},\cdots,D_{1n},D_{23},\cdots,D_{(n-1)(n)}\}$。对 D 内的距离按照从小到大的顺序进行排序，排序之后选择分位数 $k(0<k<1)$所对应的距离 D_x 作为阈值。通过分位数选择阈值，可以过滤掉击键行为中的异常值，避免干扰用户键盘行为的计算。假设用户 A 的某次输入为 $V_A=\{v_1,v_2,\cdots,v_n\}$，其中某次异常输入为 v_r，由于 v_r 的存在共产生了 $n-1$ 个异常值，分别为 $E_{r1},E_{r2},\cdots,E_{ri},\cdots,E_{r(n-1)}$。与正常击键行为的距离相比，这些距离的值更大。通过分位数可以排除距离当中的异常值带来的干扰。如当用户具有 10 次不同的击键行为且其中一次为异常值时，10 次不同的击键共会产生 45 个距离，其中异常值产生距离为 9 个，占所有距离的 20%。设置阈值对应的分位数为 $a_{0.8}$，可以排除 9 个异常值的干扰。所具有的用户的击键次数越多，由于异常值所带来的异常距离的比例会越小，通过改变分位数的大小，可以过滤掉这一部分异常距离。当距离中不存在异常值时，通过分位数设置阈值可以调节预测用户的严格程度。阈值越小，预测的条件越严格；阈值越大，预测相对宽松。

振荡幅度 $t(0<t<1)$的大小用来设定预测判断的百分比。当通过用户 A 的 n 次击键行为获得其击键行为的阈值时，对新的用户击键行为 v_x 进行预测使用如下的方法：计算 v_x 与已知集合 $V_A=\{v_1,v_2,\cdots,v_n\}$ 中所有向量的距离，得到 $s_x=\{D_{1x},D_{2x},\cdots,D_{ix},\cdots,D_{nx}\}(i=1,\cdots,n)$。设置震荡幅度为 t，当 s_x 中小于阈值的距离所占的百分比大于 t 时，接受该次击键为正常击键；当 s_x 中小于阈值的距离所占的百分比小于 t 时，认为此次击键的特征与已知击键的特征差异较大，预测该次击键为入侵者的击键。震荡幅度与分位数一样，用来控制预测时的宽松程度。t 越大，预测为正常击键的条件越严格；t 越小，预测为正常击键的条件越宽松。

3.3　与 KNN 的区别

基于距离进行键盘行为识别的方法与 KNN 十分相似，但也存在不同之处。在使用 KNN 分类时，待划分的类已确定，未知点所属的分类取决于 k 的取值。如图 3 所示，当 $k=3$ 时，未知点归为三角形分类；当 $k=5$ 时，未知点归为正方形分类。

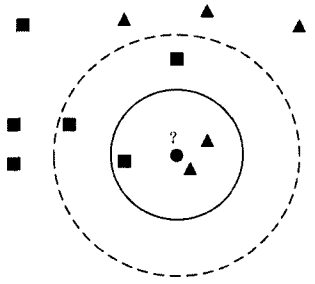

图 3　KNN 分类器示意图

而在使用距离对键盘行为进行识别时，已有数据只包含目标用户的键盘行为数据，没有负样本等其他分类数据，属于单分类问题。基于距离的键盘行为识别与 KNN 的主要区别即在于此。

4　实验部分

4.1　分位数与震荡幅度选择

在算法 1 中，如何确定分位数和震荡幅度，使得基于键盘行为的准确率最高？我们在实验中尝试了在不同分位数与不同震荡幅度的组合下，根据预测准确率的变动选择最优分位数和震荡幅度。

选择 20 个用户，每个用户的击键行为均超过 100 条。在实验中，分别设置分位数为 0.8、0.9 和 0.95；设置震荡幅度为 80%、90%和 95%。在不同的分位数和震荡幅度的组合下进行实验，预测的准确率平均值如表 1 所列。

表 1　不同分位数和震荡幅度下预测效果的对比

	80%		90%		95%	
	正	负	正	负	正	负
0.8	82.28%	98.40%	42.96%	98.20%	0.28%	98.00%
0.9	**92.66%**	**96.60%**	86.79%	96.00%	56.49%	94.80%
0.95	96.16%	92.30%	93.60%	91.60%	87.71%	90.90%

表 1 中“正”表示正样本预测准确率，“负”表示负样本预测准确率。

由表 1 可知，在震荡幅度不变的情况下，随着分位数的增大，正样本的预测准确率提高，但同时负样本的预测准确率降低。在分位数不变的情况下，随着震荡幅度的提高，正样本的预测准确率有所下降，由于震荡幅度的条件过于严格，正样本预测的准确率下降非常明显，负样本的预测准确率降低。

综合考虑正样本的预测准确率和负样本的预测准确率，确定最优分位数位为 0.9，最优震荡幅度为 80%。

4.2　向量距离选择

在本实验中通过向量之间的距离来度量击键行为之间的相似度，向量之间的距离可通过多种方式进行计算。在确定分位数和震荡幅度之后，向量距离也会影响预测的准确率。为了选择相似性度量效果最好的距离，本实验对比了 3 种距离的效果，分别为曼哈顿距离、欧氏距离和马氏距离。

在击键时间跨度超过 3 个月、击键行为数据超过 100 条的用户中，随机选择了 20 个用户进行实验。以用户某次击键的前 16 个输入构建向量。使用击键序列 $v_0=\{P_1,R_1,P_2,R_2,\cdots,P_n,R_n\}$作为向量属性，设置分位数为 0.8，震荡幅度取 80%。取每个用户击键行为中 70%的数据作为训练集用于计算阈值，剩余 30%的数据视为未知击键行为进行预测。按照算法 1 分别进行 3 组实验，每组实验所使用的数据相同。在计算向量的相似性时，第一组使用曼哈顿距离，第二组使用马氏距离，第三组使用欧氏距离。计算预测结果的准确率，得到如图 4 所示的结果。

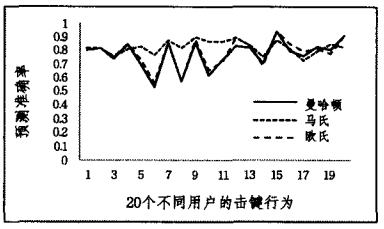

图 4　3 种不同距离的度量效果对比

图 4 表明欧氏距离和曼哈顿距离的准确率波动较大，预测准确率在 50%～80%之间波动；而马氏距离的波动较小，预测准确率稳定在 80%左右。就平均准确率而言，马氏距离的效果最好。

综上，选择马氏距离进行键盘行为的相似性计算。

4.3　向量属性选择

击键行为的相似度除了受到向量距离的影响之外，还与向量的属性相关。除了使用原生的用户击键序列之外，还可以使用用户击键的间隔时间、持续时间以及由这两个属性进行衍生的变量。本实验选择了 3 组属性不同的向量分别进行了实验，以对比不同属性在击键行为相似度度量方面的效果。

第一组向量为用户的击键序列，即从第一个击键所对应的 P 时间开始，到最后一个击键所对应的 R 为止，共 32 个时间，表示为 $b_1=\{P_1,R_1,P_2,R_2,\cdots,P_{16},R_{16}\}$。这 32 个属性直接从击键行为中提取，是未经任何衍生的原始变量。

第二组向量为击键持续时间和击键间隔时间，通过单独按键的击键持续时间和相邻按键的间隔时间构建，表示为 $b_2=\{PR_1,PR_2,\cdots,PR_{16},RP_1,RP_2,\cdots,RP_{15}\}$。

从向量 b_2 出发，可获得的衍生变量包括：PR_{avg}，表示 16 次击键持续时的平均值；RP_{avg}，表示 16 次击键中相邻两个击键之间击键间隔的平均值；PR_{std}，表示 16 次击键持续时的标准差；RP_{std} 表示 16 次击键中相邻两个击键之间击键间隔的标准差；$Time_{total}$，一次击键时的输入总时间；$Index_{RP\_max\_1}$、$Index_{RP\_max\_2}$，分别表示击键间隔时间最大的击键所出现的位置和对应折线图上的两个最大的波峰；$Index_{RP\_min\_1}$、$Index_{RP\_min\_2}$，分别表示击键间隔时间最小的击键所出现的位置和对应折线图上最小的两个波谷。

得第三组向量 $b_3=\{PR_{avg},RP_{avg},PR_{std},RP_{std},Time_{total},Index_{RP\_max\_1},Index_{RP\_max\_2},Index_{RP\_min\_1},Index_{RP\_min\_2}\}$。

选择马氏距离作为向量相似度度量方法，设置分位数为 0.8，震荡幅度取 80%，选择 3 组不同的向量属性 b_1、b_2、b_3，对 20 个不同用户的击键行为进行实验。取每个用户击键行为中 70%的数据作为训练集，剩余 30%的数据进行验证。按照算法 1 进行 3 组实验，第一组中的向量使用击键序列，第二组中的向量使用击键间隔与击键持续时间，第三组的向量使用衍生变量。使用不同向量属性时，键盘行为的预测准确率如图 5 所示。

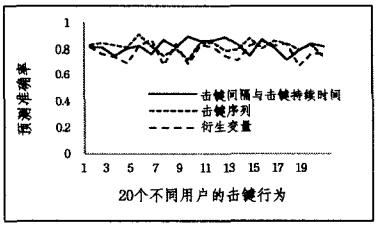

图 5　3 组不同变量的相似度度量效果

在3种向量属性中，衍生变量的波动较大，预测准确率在70%～80%之间波动。击键序列对部分用户的预测准确率达到90%，但对另外一些用户的识别准确率只有70%。击键间隔与击键持续时间对个别用户的预测准确率接近70%，但总体而言其准确率稳定在80%。

使用衍生变量进行用户识别时，由于衍生变量对击键间隔、击键持续时间进行了平均和求方差的计算，导致不同击键之间的差异信息丢失，因此在用户预测时准确率较低。击键序列是用户原生的击键顺序，虽然保留了不同用户之间的差异信息，但从第一个按键到最后一个按键之间的时间跨度很大。当计算两个向量之间的距离时，向量中较大的值可能会覆盖较小的值。如第一次击键的弹起时间 R_1 为15，第16次击键的弹起时间 R_{16} 为12000。由于最后一次击键的时间等于之前所有击键时间的累计，导致 R_{16} 在数量级上比 R_1 大很多，从而两个向量 R_1 间的差异可能被 R_{16} 间的差异所覆盖。使用击键间隔和击键持续时间进行计算时，则不存在击键序列中数量级的问题，每一个击键间隔和击键持续时间都具有相同的数量级，通过马氏距离可以将击键间隔和击键持续时间转化为相同的量纲。因此这不仅能保留不同用户之间的差异，而且不同击键之间不会存在数量级上的差异，不会出现数值较大的击键覆盖数值较小的击键的情况。

综上，选择击键间隔与击键持续时间进行用户的识别。

4.4 模拟负样本测试

正样本的测试说明，通过选择合适的分位数和震荡幅度、距离，可以很好地预测和识别出正样本，即通过该方法可以有效识别用户A自身的击键行为。但同时需要能够识别出其他用户的行为，保证不把其他用户的行为误判为用户A的行为。随机选择其他用户的击键行为进行模拟，随机选择促使负样本中具有多种击键行为，部分击键行为与用户A的击键行为较为相似，因此可能会在识别上带来干扰。这与实际中的情况是相似的，即不能确定所要预测的用户的击键行为究竟具有怎样的特征。我们选择了20个击键行为、样本数超过100的用户，并从其他用户的击键行为中随机选择了 n 条击键行为，用来模拟负样本进行检测。n 是用户A训练样本数的30%。图6是对20个不同用户击键行为进行负样本模拟后检测准确率的结果。图6表明，大多数模拟负样本的预测准确率在70%～80%之间；部分负样本的识别准确率超过了90%，原因在于此类负样本的击键行为与用户A的击键行为差异较大；同时也存在两组负样本，其识别准确率在30%～40%之间，这表明该两组负样本的击键行为与用户A的击键行为较相似，在识别时出现了较高的误判。

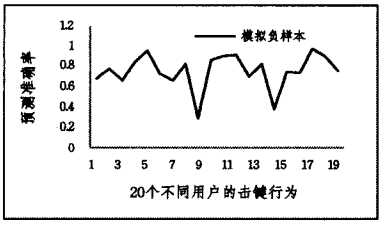

图6 使用模拟负样本的预测效果

4.5 在真实环境中的应用

在实验环境中进行键盘行为的区分性实验时，可通过预先设定的方法采集足够的正、负样本，并通过实验条件排除不同因素带来的干扰。例如，限定用户输入相同的内容，排除由于输入内容不同而带来的击键行为之间的差异；限定用户使用相同的键盘和操作系统进行输入，排除由于键盘之间的差异而带来的击键行为之间的差异。

但在真实的环境中，用户的输入内容多种多样，并且是不可控的。例如，某电商网站通过采集用户购物时输入信用卡号时的键盘行为进行用户识别，以保证交易的安全性。在这样的环境中，每个用户的信用卡号均不相同，每个用户的键盘和操作系统也各不相同。对于特定的用户而言，无法获得充足的负样本。基于分类器的训练方法由于没有足够的数据，无法进行用户识别。

本文从真实环境的击键行为出发，通过数据预处理，选择不包含不可打印字符、击键行为数据量超过100条且时间跨度超过3个月的用户击键行为进行实验。数据由于是从真实环境而非实验环境下采集的，因此可以很好地反映用户击键的行为与习惯。

使用本文的方法可以很好地解决用户正负样本不均衡和负样本缺失的情况。识别时从用户的正样本出发，通过分位数的调控选择每个用户最合适的击键行为阈值；在预测未知击键时，通过震荡幅度调节预测时的严格程度。震荡幅度越高，只有当未知击键与已知击键的相似度足够高，才能将其预测为已知用户。我们可以根据实际的需求调节震荡幅度，以控制用户识别的严格程度。

结束语 本文在真实用户击键行为数据的基础上提出了一种方法，即在用户的正负样本不均衡或负样本缺失时进行用户识别。从用户击键行为的持续时间和间隔时间等特征出发，构建向量用以描述用户的击键行为。通过实验证明，击键持续时间和击键间隔时间能获得比击键序列、击键衍生变量更高的用户预测准确率。用户击键行为之间的相似度使用向量之间的距离来度量，在计算向量时分别比较了欧氏距离、曼哈顿距离和马氏距离3种不同的距离，实验表明，使用马氏距离进行用户预测时准确率最高。用户击键行为的阈值通过分位数进行选择，计算用户击键行为之间的两两距离，将距离按升序排列，通过分位数确定阈值。在对未知击键进行预测时，计算未知击键与该用户所有历史击键之间的距离，并与阈值进行对比，通过震荡幅度控制预测时的严格程度。震荡幅度越大，预测为正常用户的条件越严格，相应未知用户与正常用户的击键行为越相似。最后，描述了用户识别方法的一个应用场景。在将来的工作中，我们将尝试对算法进行改进，以提高计算用户历史击键行为时的效率。

参 考 文 献

[1] Giot R, Hemery B, Rosenberger C. Low cost and usable multimodal biometric system based on keystroke dynamics and 2d face recognition[C]// Pattern Recognition (ICPR). Istanbul: IEEE Computer Society, 2010: 1128-1131

[2] Jain A K, Pankanti S, Prabhakar S, et al. Biometrics: a grand challenge[C]//Proceedings of the 17th International Conference Pattern Recognition, 2004. Cambridge: IEEE Computer Society, 2004: 935-942

[3] Jain A K, Ross A, Prabhakar S. An introduction to biometric recognition[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2004, 14(1): 4-20

[4] Prabhakar S, Pankanti S, Jain A K. Biometric recognition: Security and privacy concerns[J]. IEEE Security & Privacy, 2003, 1(2): 33-42

(下转第234页)

影响；c)通过加入有效的模型匹配算法更为高效准确地建立模型的引用过程。

参 考 文 献

[1] Wand Y, Weber R. Research commentary: information systems and conceptual modeling—a research agenda [J]. Information Systems Research, 2002, 13(4): 363-376

[2] Hoppenbrouwers S, Proper H A, van der Weide T P. A fundamental view on the process of conceptual modeling [M]// Conceptual Modeling ER 2005. Springer Berlin Heidelberg, 2005: 128-143

[3] Mahaux M, Nguyen L, Gotel O, et al. Collaborative creativity in requirements engineering: Analysis and practical advice [C]// 2013 IEEE Seventh International Conference on Research Challenges in Information Science (RCIS). IEEE, 2013: 1-10

[4] Theraulaz G, Bonabeau E. A brief history of stigmergy [J]. Artificial Life, 1999, 5(2): 97-116

[5] Dipple A, Raymond K, Docherty M. General theory of stigmergy: Modelling stigma semantics [J]. Cognitive Systems Research, 2014, 31: 61-92

[6] Zhang Wei, Yi Li, Zhao Hai-yan, et al. Feature-oriented stigmergy-based collaborative requirements modeling: an exploratory approach for requirements elicitation and evolution based on web-enabled collective intelligence [J]. Science China Information Sciences, 2013, 56(8): 1-18

[7] Kovitz B, Swan J. Structural stigmergy: a speculative pattern language for metaheuristics [C]// Proceedings of the 2014 Conference Companion on Genetic and Evolutionary Computation Companion. ACM, 2014: 1407-1410

[8] Ramos V, Merelo J J. Self-organized stigmergic document maps: Environment as a mechanism for context learning [J/OL]. http://urxiv.org/ftp/cs/papers/0412/0412075/pdf

[9] Rittgen P. Collaborative modeling-a design science approach [C]// Hawaii International Conference on System Sciences (HICSS). Hawaii, 2009

[10] Šmite D, Moe N B, Torkar R. Pitfalls in remote team coordination: Lessons learned from a case study [M]// Product-Focused Software Process Improvement. Springer Berlin Heidelberg, 2008: 345-359

[11] Decker B, Ras E, Rech J, et al. Wiki-based stakeholder participation in requirements engineering [J]. IEEE Software, 2007, 24(2): 28-35

[12] Xing Z, Stroulia E. UMLDiff: An Algorithm for Object-Oriented Design Differencing [C]// Proceedings of 20th IEEE International Conference on Automated Software Engineering (ASE'05). 2005: 54-65

[13] Brunet G, Chechik M, Easterbrook S, et al. A manifesto for model merging [OL]. http://folk.uio.no/shiva/research/assets/gammabo.pdf

[14] Mehra A, Grundy J, Hosking J. A generic approach to supporting diagram differencing and merging for collaborative design [C]// Proceedings of the 20th IEEE/ACM International Conference on Automated Software Engineering. ACM, 2005: 1-10

[15] Sabetzadeh M, Easterbrook S. View merging in the presence of incompleteness and inconsistency [J]. Requirements Engineering, 2006, 11(3): 174-193

(上接第 207 页)

[5] Uludag U, Pankanti S, Prabhakar S, et al. Biometric cryptosystems: issues and challenges[J]. Proceedings of the IEEE, 2004, 92(6): 948-960

[6] Vadivel A, Majumdar A K, Sural S. Performance comparison of distance metrics in content-based image retrieval applications [C]// International Conference on Information Technology. Bhubaneswar, India: CIT, 2003: 159-164

[7] Killourhy K S, Maxion R A. Comparing anomaly-detection algorithms for keystroke dynamics [C]// Dependable Systems & Networks. Lisbon: IEEE IFIP, 2009: 125-134

[8] Zhong Y, Deng Y, Jain A K. Keystroke dynamics for user authentication [C]// Computer Vision and Pattern Recognition Workshops (CVPRW). Providence, RI: IEEE Computer Society, 2012: 117-123

[9] Monrose F, Reiter M K, Wetzel S. Password hardening based on keystroke dynamics[J]. International Journal of Information Security, 2002, 1(2): 69-83

[10] Umphress D, Williams G. Identity verification through keyboard characteristics [J]. International Journal of Man-machine Studies, 1985, 23(3): 263-273

[11] Messerman A, Mustafic T, Camtepe S A, et al. Continuous and non-intrusive identity verification in real-time environments based on free-text keystroke dynamics[C]// 2011 International Joint Conference Biometrics (IJCB). Washington, DC: IEEE Biometrics Council, 2011: 1-8

[12] Al Solami E, Boyd C, Clark A, et al. Continuous Biometric Authentication: Can It Be More Practical? [C]// 2010 12th IEEE International Conference High Performance Computing and Communications (HPCC). Melbourne, VIC, 2010: 647-652

[13] Zack R S, Tappert C C, Cha S H. Performance of a long-text-input keystroke biometric authentication system using an improved k-nearest-neighbor classification method [C] // 2010 Fourth IEEE International Conference, Biometrics: Theory Applications and Systems (BTAS). Washington, DC: IEEE Biometrics Council, 2010: 1-6

[14] Sim T, Janakiraman R. Are digraphs good for free-text keystroke dynamics? [C]// Computer Vision and Pattern Recognition, 2007(CVPR). Minneapolis, MN, 2007: 1-6

[15] De Ru W G, Eloff J H P. Enhanced password authentication through fuzzy logic[J]. IEEE Expert, 1997, 12(6): 38-45

[16] Tey C M, Gupta P, Gao D. I can be You: Questioning the use of Keystroke Dynamics as Biometrics[C]// The 20th Annual Network & Distributed System Security Symposium (NDSS). San Diego, CA, 2013: 134-150

[17] Cho S, Han C, Han D H, et al. Web-based keystroke dynamics identity verification using neural network[J]. Journal of Organizational Computing and Electronic Commerce, 2000, 10(4): 295-307

[18] Joyce R, Gupta G. Identity authentication based on keystroke latencies[J]. Communications of the ACM, 1990, 33(2): 168-176

[19] Kang P, Hwang S, Cho S. Continual retraining of keystroke dynamics based authenticator [M] // Advances in Biometrics. Seoul, Korea: Springer Berlin Heidelberg, 2007: 1203-1211

[20] Yu E, Cho S. GA-SVM wrapper approach for feature subset selection in keystroke dynamics identity verification[C]// Proceedings of the International Joint Conference Neural Networks, 2003. Portland, OR, 2003: 2253-2257