计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2017,26(11):176−181 [doi: 10.15888/j.cnki.csa.006056]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于 SVM 算法的用户行为认证方法①

程建峰[1], 乐　俊[2], 刘　丹[1]

[1](电子科技大学 电子科学技术研究院, 成都 611731)
[2](西南电子信息技术研究所, 成都 610041)

摘　要: 为提高手机安全性, 提出一种基于 SVM 的用户操作行为认证方法. 通过监听手机触摸屏设备, 持续获取用户操作时的滑动轨迹、接触面积等原始数据. 设计用户行为特征提取算法以建立用户特征样本, 经 SVM 算法加以训练形成用户行为特征模型; 综合用户访问目标及历史认证结果采用不同认证策略, 达到重点保护敏感数据, 方便用户访问非敏感数据的效果. 在 Android 系统环境下的实验验证表明, 该方法具有良好的认证效果.

关键词: 行为分析; 持续认证; SVM 算法; Android 系统

引用格式: 程建峰,乐俊,刘丹.基于 SVM 算法的用户行为认证方法.计算机系统应用,2017,26(11):176–181. http://www.c-s-a.org.cn/1003-3254/6056.html

User Behavior Authentication Method Based on SVM Algorithm

CHENG Jian-Feng[1], YUE Jun[2], LIU Dan[1]

[1](Research Institute Electronic Science and Technology of UESTC, University of Electronic Science and Technology, Chengdu 611731, China)
[2](Southwest Electronic Information Technology Research Institute, Chengdu 610041, China)

Abstract: In order to enhance security of the mobile phone, a method of user operation behavior authentication based on SVM is proposed. By monitoring the mobile phone touch screen device, continuous access to user operation when the sliding track, contact area and other raw data. The user behavior feature extraction algorithm is designed to establish the user characteristic sample, and the SVM algorithm is used to train the user behavior characteristic model. The user access goals and historical certification results are integrated with different authentication strategies to achieve key protection sensitive data to facilitate the user access to non-sensitive data. Experimental verification in the Android system environment shows that the method has a good authentication effect.

Key words: behavior analysis; continuous certification; SVM algorithm; Android system

智能手机的广泛应用极大方便了人们的生产和生活, 接踵而至的安全问题却给人们的隐私及财产安全带来极大隐患. 传统的安全研究主要集中在抗病毒、防木马等系统防护方向. 一项研究表明 56% 的智能手机用户为使用方便, 使用极简单的密码或干脆不使用密码[1], 而这使一些不法分子有了可乘之机. 为保护用户财产隐私安全, 研究工作者提出了生物信息认证技术与行为信息认证技术[2]. 生物信息认证技术主要包括: 指纹识别技术; 语音识别技术; 图像识别技术等[3]; 但这些技术本身需要特殊硬件支持且易受外界环境干扰, 且生物信息特征明显、易于提取, 不法分子可以通过伪造、提取用户信息绕过认证系统.

目前, 行为认证方式研究相对较少. Saurabh Singh 等人通过实验说明了用户行为特征的相对稳定性[4], 解释了使用用户行为特征作为认证依据的合理性. Oriana Riva 等人提出, 通过采集加速传感器数据分析用户步

① 收稿时间: 2017-02-28; 修改时间: 2017-03-16; 采用时间: 2017-03-20

态特征, 确定手机始终是处于合法用户的控制下, 从而达到减少一些不必要认证的目的[5]. Chao Shen 等人通过采集用户预定义滑动轨迹数据说明了行为认证的有效性[6]. 以上研究说明了行为认证技术的可行性, 但以上成果应用范围较为局限: Oriana Riva 等人方案特征数据单一, 仅适用于经常移动的用户; Chao Shen 等人方案要求用户输入固定轨迹且没有摆脱一次认证持续授权的弊端.

为解决以上弊端, 文章提出一种通过分析整理用户操作信息、用户操作上下文, 学习用户行为习惯的方法, 从而实现一种持续认证系统. 该系统以后台服务形式运行在宿主机上, 收集用户在正常使用时所产生的操作数据, 操作数据主要包括: 滑动轨迹、指尖大小、触摸压力. 这样不需要用户刻意配合特征提取工作, 即可完成用户特征建立及用户认证授权, 在提升系统安全的同时不给用户正常使用添加额外负担.

1 系统设计

认证试验系统采用模块化设计方式, 以后台服务形式运行在 Android 系统平台. 其系统框架如图 1 所示.

图 1 系统总体框图

数据采集模块主要监听手机输入设备文件, 采集、整理设备数据形成滑动、触摸面积等操作数据. 数据特征提取模块则通过分类、计算、过滤加工生成符合 SVM(Support Vector Machine) 算法需要的特征样本数据. 不同样本数据生成相应的分类器, 分类器具备检验数据是否符合用户行为特征的能力. 认证判决模快综合用户访问对象及认证结果决定是否阻止用户操作决定.

认证系统运行分为训练和认证两大阶段. 在系统训练阶段: 系统不断收集、整理、分类用户操作数据以建立相应的用户样本集, 通过 SVM 算法对样本数据加以训练生成用户行为特征模型. 在认证阶段: 分析待检验用户操作数据, 判断操作数据类型, 找到该数据类型对应的用户行为特征模型, 以该特征模型检验数据合法性, 并输出检验结果. 系统工作流程如图 2 所示.

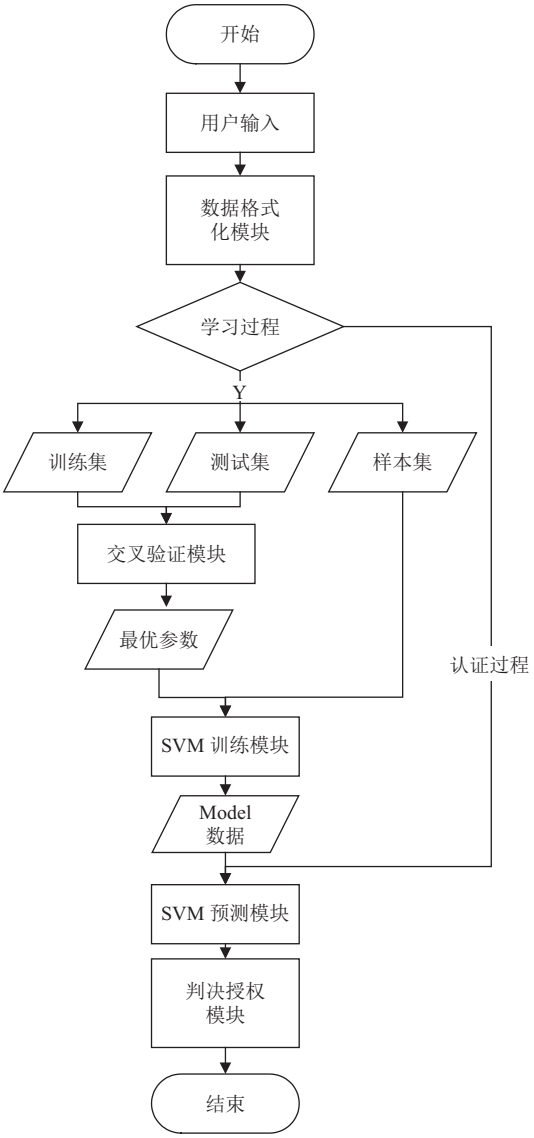

图 2 系统工作流程

2 分类器设计

分类器的好坏对于用户行为特征模型建立至关重要, 而分类器的质量取决于是否找到合适的分类算法. SVM 作为数据挖掘的一种新方法[7], 成功的解决了小样本数据的预测和分类问题, 且 SVM 算法具备训练速

度快的优点, 非常适合在运算速度有限的手机上运行.

2.1 分类器关键算法

SVM 算法核心思想是将样本集合映射到高维空间, 并在高维空间寻找一个最优超平面分割样本, 同时使正负样本超平面间的几何间隔最大. SVM 算法为简化算法复杂度引入了核函数概念, 常用的核函数有: Linear 核函数和 RBF 核函数. 其中 RBF 核函数常在线性不可分情况下使用, 如式 (1) 所示:

$$K(x,x_i)=\exp(\frac{-\|x-x_i\|^2}{2\sigma^2}) \tag{1}$$

其中, $2\sigma^2$ 含义为核函数半径 g, 核函数半径对分类器预测精度有较大影响, 通常通过交叉验证方式确定最优参数. 在实际应用中, 样本数据中不可避免的被噪声数据污染, 这些数据对分类结果可能造成较大影响. 为此 SVM 算法引入惩罚因子 c 防止过度拟合, c 参数大小反映了对离群点的重视程度. 其数学模型如式 (2) 所示:

$$\begin{aligned}&\min_{r,w,b}=\frac{1}{2}\|w\|^2+c\sum_{i=1}^{m}\xi_i\\&s.t.\ y^{(i)}(w^Tx^{(i)}+b)\geq 1-\xi_i\geq 0, i=[1,m]\end{aligned} \tag{2}$$

令:

$$g_i(w)=-y^{(i)}(w^Tx^{(i)}+b)+1-\xi_i\leq 0 \tag{3}$$

不等式变形为:

$$s.t.\ g_i(w)<0,\quad \xi_i\geq 0, i=[i,m] \tag{4}$$

求解式 (4), 得判别公式如下:

$$\begin{aligned}f(x,a)&=sgn\{\sum_{x_i}^{l}a_iy_iK(x_i,y_j)-b\}\\&=\begin{cases}1, & \sum\limits_{x_i}^{l}a_iy_iK(x_i,x_j)-b\geq 1\\0, & \sum\limits_{x_i}^{l}a_iy_iK(x_i,x_j)-b\leq -1\end{cases}\end{aligned} \tag{5}$$

上述判别式即可用于构建用户预测模型.

2.2 分类器参数寻优

SVM 算法发展较为成熟, 本实验系统使用了 LibSVM 算法工具箱中的 java 模块[8], 该模块已经完成了 SVM 算法实现工作. 该模块涵盖的应用场景较为全面, 在使用时还需要根据实际用途选择合适的应用场景.

行为认证技术主要依据用户样本数据建立用户预测模型, 其预测模型只需判断测试数据是否符合用户模型, 这是典型的一分类案例, 因此本次实验主要应用 SVM 一分类算法. svm 一分类算法实质上只是二分类算法的一种特殊情况, 即只有正类 (目标样本) 样本数据或极少量负类 (非目标对象样本) 样本, 因此一分类算法引入参数 n 表明样本数据中负类样本占总样本容量的百分比. SVM 算法不同的核函数对预测精度有较大影响, 需要通过实验对比两种核函数建立模型的预测精度, 确定合适的核函数. 实验在其他参数为最优的情况下, 预测精度随参数 n 发生如图 3 所示变化.

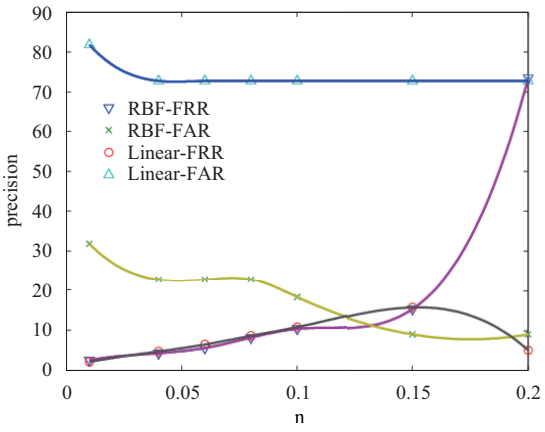

图 3 预测精度

通过图 3 发现, linear 核函数的 FAR(False Acceptance Rate) 大大高于 RBF 核函数的 FAR, 而 FRR(False Rejection Rate) 两种核函数总体差距不大. 这是因为用户模型建立所需数据种类繁多, 很难通过线性关系来描述, 因此系统中 SVM 算法模块使用 RBF 高斯核函数.

在实验中, 我们只采集用户特征数据, 因此训练样本中不存在负类样本, 此时参数 n 可以理解为: 将用户特征空间一定比例的边缘数据认为是负类样本, 在这种情形下, 离群点的惩罚因子 c 大小便不再重要, 因为参数 n 的存在, 已经将离群点认为是负类样本. 目前, 还没有通过理论计算确定最优参数 n、g 的方法. 系统采用较为常用的网格搜索法, 根据经验参数 g 与训练数据维度有关, 实验主要在维度倒数附近搜索, 而 n 参数也在 0.1 附近搜索, 图 4 和 5 分别为不同参数下的特征模型的 FAR 和 FRR 结果.

通过对比两幅图我们发现随着 FRR 升高 FAR 逐步降低, 系统中最优参数确定不能根据最优 FRR 或 FAR 确定, 系统中采用了保障 FAR 和 FRR 低于 20% 的情况下取 FAR+FRR 的最低值, 这组 (n, g) 作为最优参数.

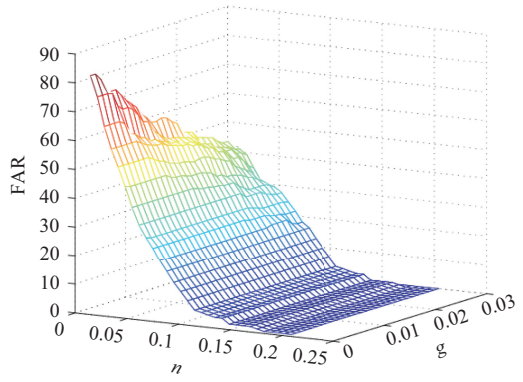

图 4 FAR 分布图

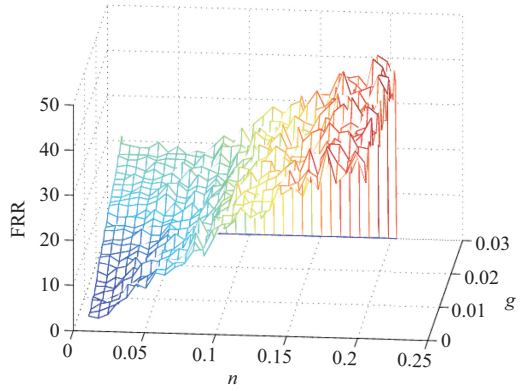

图 5 FRR 分布图

3 特征数据提取

用户特征模型建立, 离不开用户特征数据样本库. 在用户使用过程中, 收集到的用户数据种类繁多, 在本次实验中主要采集用户在触摸屏上操作时产生的滑动轨迹、点击事件、触摸面积、时间戳等数据 (因设备限制未能采集到触摸压力这一重要特征). 用户在触摸屏上滑动时, 系统将捕捉到一系列 P(x, y) 坐标点阵, 在每一个坐标 P(x, y) 触发时获取当时触摸面积 s、时间戳 t 形成数据 C(x, y, s, t), 这样用户一次滑动操作即可得到元素为 C(x, y, s, t) 的一维向量 V[c_1, c_2…c_n]. 随着用户不断的使用便可获取足够的样本数据.

上述数据为用户原始数据, 每一维向量都与用户滑动起始位置相关, 为不限制用户使用, 特征模型应与用户操作位置无关. 为此通过相邻两个元素时间与位置变化计算出用户滑动斜率 k 与速度大小 v 变化形成带测样本数据元素 S(k, v, s). 通过以上方法处理后形成样本数据为: T(滑动总时间)、S_1(k, v, s)… Sn(k, v, s). 以上样本数据建立的用户特征模型预测精度低, 其 FRR 与 FAR 在 50% 左右浮动. 通过观察对比发现: 数据序列中包括大量重复数据, 这些数据增加了训练维度但对预测精度毫无帮助, 且样本数据向量维数差异较大, 滑动操作方向混杂. 过多的操作类型使用户样本数据特征繁多, SVM 算法所求超平面不得不囊括所有特征, 这将使所求超平面过大, 使特征模型因过度学习导致预测精度下降.

为克服以上弊端, 设计过滤模块, 依据原始数据 C(x, y, s, t) 中 x, y 值合并相邻重复数据, 合并完成后一维向量中不存在相同 x, y 元素. 同时实验从滑动距离、滑动方向入手, 将用户数据分类. 实验时将滑动方向分别分为两、四、八个方向, 发现四方向分类预测精度比两方向高, 比八方向分类预测精度提高并不明显. 为减少分类器数量, 本实验采用四方向分类: 右上; 右下; 左下; 左上. 滑动距离按数据长度分为五类, 如表 1 所示.

表 1 滑动距离按数据长度分类

A	B	C	D	E
[1, 5]	[6, 10]	[11, 15]	[16, 20]	[21, 25]

通过以上分类, 样本集组合达 20 种, 实验中分别对这 20 种数据样本训练, 为用户生成不同模型. 上述数据生成后数据值范围分布不均, 需要进行归一化处理, 处理后以某一维数据为例, 其样本数据格式如表 2 所示.

表 2 某一维数据的样本数据格式

总时间(ms)	斜率y/x	速度(px/ms)	指尖面积 (px*px)
1:398	2:-12.0	3:0.25086	4:0.62

每一次触摸事件对应生成斜率、速度、面积数据, 因此样本总维数为: 总时间+3*维数, 其范围为[2, 76], 样本文件按类分为为 20 个. 到此特征数据获取工作完成.

4 认证判决

鉴于用户行为习惯的相对稳定性, 一旦用户预测模型建立完成后, 其近期的操作总体符合该模型, 因此预测模型的预测结果可以作为判别用户真实身份的有效依据. 实验中预测模块的精度并不能达到 100% 正确, 其中用户本人操作时 76.8%~88.8% 的操作符合预测模块, 其他用户操作时有 2.7%~26% 的操作符合用户操作, 因此将一次操作结果作为认证依据会导致认证错误率很高. 从统计角度看, 用户操作合法率一般会趋于某个值, 基于此思想判决模块将记录用户操作历

史结果, 综合分析用户历史操作及当前操作做出判断. 实际使用中用户一次任务往往需要多次操作才能完成, 这也使得基于统计方法减少错误认证的概率成为可能. 实际使用时用户访问浏览器、地图与访问图库、支付程序时, 其敏感程度大不相同, 因此认证判决时其策略也应不同. 对于非敏感访问, 认证策略为: 统计用户连续五次 (不足五次的以实际次数为准) 操作判决结果的平均值, 若五次操作中 4 次及以上为负或连续 3 次操作判断为负则拒绝认证. 以上述预测精度为例, 则错误拒绝访问的概率为 0.45%, 错误接受访问的概率为 6%. 对于敏感访问, 认证策略为: 统计用户连续五次 (不足五次的以实际次数为准) 操作判决结果的平均值, 若五次操作中 3 次及以上为负操作判断为负则拒绝认证. 以上述预测精度为例, 则错误拒绝访问的概率为 4.3%, 错误接受访问的概率为 0.85%. 基于操作的敏感程度可以更改认证阀值, 调和错误拒绝访问与错误接受访问的矛盾.

5 实验及分析

5.1 环境

行为认证系统运行在 Android 系统用户态. 由于 Android 安全机制限制, 认证试验系统无法通过系统 API 跨平台获取用户操作数据, 因此认证实验系统以 root 权限运行在 Android 用户态, 通过监听系统输入设备文件跨进程获取用户操作数据.

本次实验硬件环境为魅族 U10, 操作系统为 Android 5.1.1. 实验所采用 APP 数据模块根据上文数据特征提取要求自行实现, 分类器及预测模块使用了 LIVSVM 库中公用 API.

5.2 数据来源

实验数据采集自二十名同学, 其中七位女同学和十三位男同学. 二十名同学在使用时主要进行滑动操作, 并在两天的不同时刻执行上述操作. 应用程序收集同学操作数据, 数据经过滤、分类、特征计算、归一化处理后生成相应特征样本数据. 数据总量在 5000 条左右, 其中长度为 B 和 C 的数据比例较大, 约占总量的 60%. 随机样本数据分为训练集 $S_i(1\le i\le 6)$ 和测试集 $T_i(1\le i\le 6)$, 其中训练集占数据总量 70%, 测试集为 30%.

5.3 实验结果

通过搜索法确定最优核函数半径后 g, 不同的参数 n 得出实验结果如表 3 所示.

表 3 实验结果 (单位: %)

N	FRR	RAR
0.01	11.2	23.6
0.02	16.0	13.7
0.03	18.4	10.1
0.04	21.6	7.6
0.05	23.2	4.5

5.4 实验分析

分类器根据用户指尖大小、速度变化、斜率变化将用户特征映射到高维空间中某一超平面. FAR 4.5%~23.6% 说明不同用户间的超平面仍有重叠, 而 FRR 11.2%~23.2% 说明用户超平面未能覆盖所有用户的特征. 不同的参数 n, 意味着训练集中不同的比例代表用户特征, 随着 n 变大, 用户特征超平面随着缩小, 其 FAR 也随即降低, 但 FRR 升高, 这说明上述特征还不足以区分不同的身份, 还需要提取更多的特征. 在实验中发现, 与训练数据产生时间相近的数据, 其预测精度较高, 随着时间推移预测精度呈逐渐降低的趋势, 即用户行为特征是相对稳定的, 因此分类器设计需要考虑增量学习模块, 限于篇幅这里不再论述.

6 结束语

当前, 基于密码、指纹、语音的认证技术已经较为成熟, 但每种技术都有其自身的局限性, 且都没有摆脱一次认证持续授权的模式. 基于行为认证技术不需硬件支持、不易被模仿、相对稳定且可以在用户使用中持续监视认证的特性, 可以作为传统认证方式的一种有效补充. 本系统中因硬件限制, 没有采集像压力等较明显的特征数据, 这使得系统预测精度较低. 随着手机中更多传感器的应用, 用户特征数据也将更加丰富, 其系统预测也将更加准成. 因此我们可以相信基于用户行为认证技术也将有更广泛的应用.

参考文献

1 Khan H, Atwater A, Hengartner U. Itus: An implicit authentication framework for android. Proc. of the 20th Annual International Conference on Mobile Computing and Networking. Maui, Hawaii, USA. 2014. 507–518.

2 Tanviruzzaman M, Ahamed SI. Your phone knows you: Almost transparent authentication for smartphones. Proc. of the 38th Annual Computer Software and Applications Conference (COMPSAC). Vasteras, Sweden. 2014. 374–383.

3 郑方, 艾斯卡尔•肉孜, 王仁宇, 等. 生物特征识别技术综述. 信息安全研究, 2016, 2(1): 12–26.

4 Singh S, Sinha M. Pattern construction by extracting user specific features in keystroke authentication system. Proc. of the 4th International Conference on Computer and Communication Technology (ICCCT). Allahabad, India. 2013. 181–184.

5 Riva O, Qin C, Strauss K, *et al*. progressive authentication: Deciding when to authenticate on mobile phones. Proc. of the 21st USENIX Conference on Security Symposium. Bellevue, WA, USA. 2012. 15.

6 Shen C, Zhang Y, Cai ZM, *et al*. Touch-interaction behavior for continuous user authentication on smartphones. Proc. of the 2015 International Conference on Biometrics (ICB). Phuket, Thailand. 2015. 157–162.

7 汪海燕, 黎建辉, 杨风雷. 支持向量机理论及算法研究综述. 计算机应用研究, 2014, 31(5): 1281–1286.

8 Chang CC, Lin CJ. LIBSVM: A library for support vector machines. ACM Transactions on Intelligent Systems and Technology (TIST), 2011, 2(3): 27.