计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.200600004

基于上采样单分类的智能手机手势密码隐式身份认证机制

姚沐言　陶　丹
北京交通大学电子信息工程学院　北京 100044
(20120019@bjtu.edu.cn)

摘　要　现有智能手机往往使用广泛且存储有敏感信息,一旦丢失会造成巨大的安全隐患,故数据安全的重要性日益凸显。鉴于传统认证策略的脆弱性,提出了一种基于上采样单分类的隐式身份认证机制。首先,融合使用了时间、二维及三维等多类手机内置传感器从不同维度采集用户的行为特征。其次,为降低高维数据所含噪声对分类的影响,提出了一种精选特征并降维的行为特征筛选方法,对所提取的特征进行向量排序、筛选以及降维。特别地,考虑到现有基于二分类算法方案的局限性,采用SVM SMOTE对正样本数据进行上采样,并提出了基于单分类的认证决策机制,以在单类小规模训练集上实现分类。最后基于实际的样本集进行性能测试,结果表明,所提方案在准确率、FAR、FRR与AUC指标上的表现部分优于使用大规模数据进行训练的传统KNN二分类器。

关键词:隐式身份认证;手势密码;单类支持向量机;超小规模训练集;上采样
中图法分类号　TP391.4

Implicit Authentication Mechanism of Pattern Unlock Based on Over-sampling and One-class Classification for Smartphones

YAO Mu-yan and TAO Dan
School of Electronic and Information Engineering, Beijing Jiaotong University, Beijing 100044, China

Abstract　Nowadays, smartphones are widely used and stored with sensitive information, and the loss of any personal device can cause fatal information compromise. Thus, the people's attention towards data security has been elevated to a higher level. Considering the delicacy of traditional authentications, this paper investigates an implicit authentication mechanism based on over-sampling and one-class classification, for pattern unlock on smartphones. First, a fusion of time, two-dimensional and three-dimensional sensors is employed, to collect user behavioral biometrics comprehensively. Second, in order to ease the impact caused by noise contained in high-dimensional data, a feature screening, which is composed of feature selection and dimensional compression, is designed. Particularly, in view of the existing limitations of the current binary classification schemes, SVM SMOTE is used to over-sample the user behavioral data, and a one-class classification authentication mechanism is delivered to implement classification, of which the learning process is only based on a single-class diminutive training set. A series of experiments have been conducted on actual data, and results show that the proposed scheme, when only relies on a single-class diminutive training set, performs partially better than the traditional binomial KNN classifier which is trained on large-scale data, in terms of accuracy, FAR, FRR and AUC.

Keywords　Implicit authentication, Gesture pattern, One-class support vector machine, Diminutive training data set, Over-sampling

1　引言

智能手机在人们日常生活中承载的功能逐渐多样化,隐私与数据安全的重要性亦日益凸显。传统的显式认证方案不会分析用户解锁时的行为特征,因此对于肩窥攻击、口令穷举、污迹攻击在内的多种认证攻击方法无能为力[1-2]。作为改进,引入生物识别技术后的鉴权方案要求被访问的设备在检测信令正确与否的同时,一并使用各种传感器核对访问者的身份。此过程对用户而言不可感知,因此被称为隐式身份认证。

以往的隐式认证方案仅依靠单类传感器[3],信息有限,实际应用时准确率较低。随着智能手机应用场景的不断变化,

到稿日期:2020-06-01　返修日期:2020-09-14　本文已加入开放科学计划(OSID),请扫描上方二维码获取补充信息。
基金项目:国家自然科学基金"面上"基金项目(61872027);综合业务网理论及关键技术国家重点实验室开放研究基金(ISN21-16)
This work was supported by the National Natural Science Foundation of China (61872027) and Open Research Fund of State Key Laboratory of Integrated Services Networks (ISN21-16).
通信作者:陶丹(dtao@bjtu.edu.cn)

对身份认证算法在各类场景下的适用性也提出了更高的要求。现有研究工作往往倾向于使用二分类器完成身份认证的操作[4-6]，受原理所限，这类方案需要分布均衡且预先标注至少两类样本的数据来训练模型，难以实际应用。考虑到二分类方案的限制，有学者提出了基于单分类器的方案，例如文献[7-8]的研究基于单类支持向量机来解决 PIN 密码的解锁问题；文献[9]对 APP 内的滑动特征进行建模。然而，此类方案仍然需要录入大量的机主数据以实现较好的分类性能。

在此基础上，为了解决用户初始录入数据量的限制，本文提出了一种以用户使用手势密码的行为特征为基础的，基于上采样、单分类的隐式身份认证机制。该机制通过将智能手机用户的划屏行为与握持姿势作为可用特征，通过对分类器进行训练来实现身份认证的目的。特别地，考虑到在录入数据阶段的用户体验，本方案使用了特征筛选、数据扩增、单分类的方法以在较少的初始数据下仍能保证分类性能。

2 系统框架

基于上采样单分类的智能手机手势密码隐式身份认证的框架如图 1 所示。

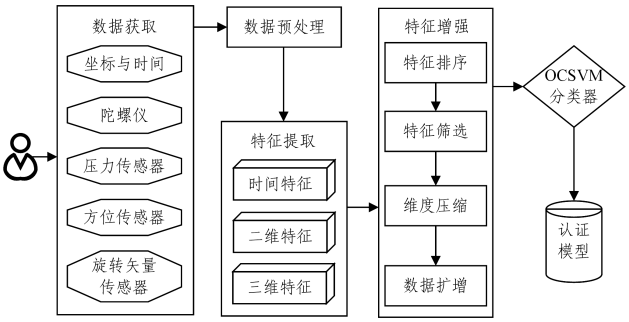

图 1 基于上采样单分类处理的隐式身份认证框图

Fig. 1 Framework diagram of implicit authentication based on over-sampling and one-class classifier

在用户与设备进行交互的同时，设备利用屏幕内置的触摸坐标传感器、压力传感器、时间计数器、陀螺仪、方位传感器与旋转矢量传感器对用户行为数据进行采样，从而对 3 类不同粒度的特征信息进行提取。随后，本方案对特征进行重要性排序，并筛选出最佳的特征组合。为了有效减弱高维数据所含噪声对分类造成的影响，对数据维度应用降维的操作。最后，使用由网格搜索得到的最优参数对单类支持向量机(One-class SVM)进行训练，从而得到认证模型，并利用此模型对来自未知用户的行为数据进行匹配，以判定用户身份。

3 数据与特征提取

本文使用实验室自建数据集，其含有 12600 条原始数据，采集自 42 名参与者(22 名男性与 20 名女性)。

在预设密码时，基于常用的手势密码设定原则[10]，纳入对密码所用点数、重叠程度、输入难易的考量，在尽量保证所选手势密码具有代表性的前提下筛选出 4 组手势密码(见图 2)以评估所提方案的有效性、普适性。其中 L，Z，S 型手势密码所用点数分别为 5 点、7 点、9 点，涵盖了点数全选与未全选的情况，对应的平均输入时长与图形转折次数也不断增加；而 T 型手势密码存在录入重叠的现象，存在视觉复杂性。

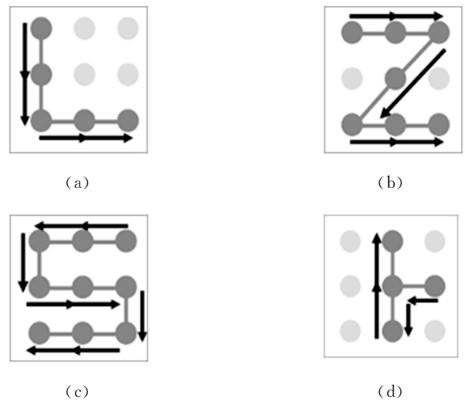

图 2 所用手势密码的样例

Fig. 2 Samples for involved patterns

为获取用户的滑动行为数据，采集智能手机的 6 种不同的数据用于表征用户的行为，分别为坐标数据、时间计数器、陀螺仪、压力传感器、方位传感器与旋转矢量传感器。其中，用户解锁的时长、坐标信息可以记录用户的滑动偏好；陀螺仪传感器数据包括用户解锁过程中较大程度的运动幅度数据；方位传感器数据可以反映用户解锁过程中的较细微旋转、倾斜等情况；旋转矢量传感器数据包含解锁时的手持设备空间朝向与抖动情况。

图 3 展现了不同用户在测试设备上重复做出 20 次含有 5 个解锁点的 L 形手势(见图 2(a))时，在每个解锁点上所耗时间的平均分布(记绘制完成时为 100%)，即做出手势密码图形时的绘制节律。由图 3 可知，即便是在绘制相同的手势密码图形时，不同用户的行为特征亦存在较大的差异。合理利用这些数据，可以对用户的身份加以区分。

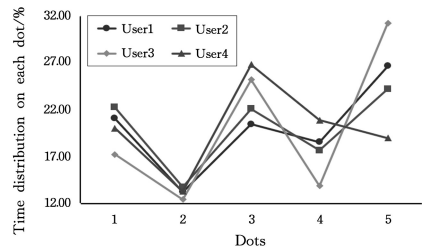

图 3 不同用户做出 L 形手势的绘制节律

Fig. 3 Time distribution on each dot of different users drawing L pattern

基于用户行为特征动态数据信息，本文从时长、二维与三维 3 个维度构建用户的行为特征，默认当前手势的点数为 n。

(1)时长特征

时长特征指时长相关特征，规定向量 $\boldsymbol{T}_d=\{t_{d1},t_{d2},\cdots,t_{dn}\}$，$\boldsymbol{T}_c=\{t_{c1},t_{c2},\cdots,t_{cn}\}$ 分别为长度为 n、以时基 $\boldsymbol{T}=\{T_1,T_2,\cdots T_n\}$ 进行 n 次采样得到的用户在每一解锁点上的停留时间 t_{dwell}，以及用户到达每一解锁点时的累计耗时 t_{cum} 的序列。

解锁过程中受肌肉记忆的影响，不同用户在绘制相同手势密码时的节律存在差异，耗时分布具有特异性，因此本部分特征所含有的信息能够有效地对用户的身份进行区分。

(2)二维特征

二维特征指用户解锁过程中，手指与被访问设备显示屏交互时所产生的行为数据。

本文利用屏幕返回的触摸坐标对用户解锁过程中的滑动轨迹进行刻画。使向量 $\boldsymbol{X}=\{x_1,x_2,\cdots,x_k\}$，$\boldsymbol{Y}=\{y_1,y_2,\cdots,y_k\}$ 分别为以时基 $\boldsymbol{T}=\{T_1,T_2,\cdots T_k\}$ 进行 k 次采样得到的屏幕坐标系中 x 和 y 坐标的数据，其中采样次数 $k=t_{cum}/10\,\mathrm{ms}$。由此可以推算得到，用户同每一解锁点交互时的总移动距离 $d_i=\sum_{m=2}^{m=k}\left(\sqrt{(x_m-x_{m-1})^2+(y_m-y_{m-1})^2}\right)$，$i\in[1,n]$，从而有长度为 n 的数据向量 $\boldsymbol{D}_{cum}=\{d_1,d_2,\cdots,d_n\}$。

此外，设备使用屏幕内嵌的压力传感器记录用户滑动过程中对各触摸点施加的压力。具体地，使 $\boldsymbol{P}_{\max}=\{p_{m1},p_{m2},\cdots,p_{mn}\}$，$\boldsymbol{P}_{\mathrm{avg}}=\{p_{a1},p_{a2},\cdots,p_{an}\}$，$\boldsymbol{P}_{\min}=\{p_{i1},p_{i2},\cdots,p_{in}\}$ 分别为长度为 n、以时基 $\boldsymbol{T}=\{T_1,T_2,\cdots T_n\}$ 进行 n 次采样的用户输入的屏幕施加压力的最大、平均、最小值 $\boldsymbol{p}_{\max}$，$\boldsymbol{p}_{\mathrm{avg}}$，$\boldsymbol{p}_{\min}$ 向量数据。相似地，构建长度为 n、以时基 $\boldsymbol{T}=\{T_1,T_2,\cdots T_n\}$ 进行 n 次采样的用户交互过程中与所显示的解锁点接触的起始、结束时刻的屏幕坐标系坐标向量 $\boldsymbol{X}_{\mathrm{start}}$，$\boldsymbol{X}_{\mathrm{end}}$，$\boldsymbol{Y}_{\mathrm{start}}$，$\boldsymbol{Y}_{\mathrm{end}}$。

在触发由相同按钮决定的事件(event)时，不同的用户点击按钮的具体部位存在细微差异。此部分的特征可以反映出用户在进行解锁操作过程中的点击行为与滑动行为偏好。

(3)三维特征

三维特征指用户尝试解锁设备时，被访问设备由于用户相关肌肉发生牵连而造成的设备姿态在三维空间中发生改变的特征。具体地，在解锁过程中，被握持设备会随用户手掌及腕部的晃动发生三轴速度、加速度、重心、仰角、方位角、旋转速率等参数变化的情况。

与屏幕施加压力向量数据的构造方法相同，构造长度均为 n、以时基 $\boldsymbol{T}=\{T_1,T_2,\cdots T_n\}$ 进行 n 次采样的陀螺仪三轴数据 $\boldsymbol{Gyr}_x$，$\boldsymbol{Gyr}_y$，$\boldsymbol{Gyr}_z$ 向量，方位传感器三轴数据 $\boldsymbol{Ori}_{azimuth}$，$\boldsymbol{Ori}_{pitch}$，$\boldsymbol{Ori}_{roll}$ 向量，旋转矢量传感器的三轴数据 $\boldsymbol{RV}_x$，$\boldsymbol{RV}_y$，$\boldsymbol{RV}_z$ 向量。延伸地，对于 $i\in[1,n]$，有旋转矢量传感器的标量数据 $\boldsymbol{RV}_{si}=\sqrt{RV_{xi}^2+RV_{yi}^2+RV_{zi}^2}$，从而得到长度为 n 的向量 $\boldsymbol{RV}_{scalar}=\{RV_{s1},RV_{s2},\cdots,RV_{sn}\}$。

在完成前述的特征提取操作后，构建长度为 $20n$ 的用户行为数据特征向量 $\boldsymbol{F}=\{\boldsymbol{T}_d\parallel\boldsymbol{T}_c\parallel\boldsymbol{D}_{\mathrm{cum}}\parallel\boldsymbol{P}_{\max}\parallel\boldsymbol{P}_{\mathrm{avg}}\parallel\cdots\boldsymbol{RV}_z\parallel\boldsymbol{RV}_{\mathrm{scalar}}\}=\{p_1^1,p_2^1,p_3^1,\cdots,p_n^1,p_1^2,p_2^2,\cdots,p_n^2,\cdots,p_n^{20}\}$，以对用户的行为模式进行进一步分析。

4 多级特征处理与身份认证

4.1 特征排序与筛选

由于用户行为数据中不同特征对分类结果的影响程度不同，且为了减小不同传感器对同一事件进行记录时潜在的冗余信息（数据特征不同分量之间可能相关）对分类结果造成的影响，需要对用户行为数据中特征分量的重要性进行排序。本文选择基于 XGBoost 的方案[11]对此进行实现。

步骤 1　对于每一条用户行为数据特征向量，在提升树被创建后，由梯度提升算法直接得到每个分量的重要性得分。

步骤 2　以步骤 1 中的重要性得分为依据，对 $\boldsymbol{F}=\{p_1^1,p_2^1,p_3^1,\cdots,p_n^1,p_1^2,p_2^2,\cdots,p_n^2,\cdots,p_n^{20}\}$ 进行降序重排列，可以得到特征分量顺序改变的用户行为特征向量 $\boldsymbol{F}'=\{p_1',p_2',p_3',\cdots,p_{21}',\cdots,p_{20n}'\}$。

步骤 3　对用户行为特征向量 $\boldsymbol{F}'$ 的子集 $\boldsymbol{P}\subseteq\boldsymbol{F}'$，$P\neq\varnothing$ 进行遍历，对每一个遍历组合使用单类支持向量机计算 ROC(Receiver Operating Characteristic)曲线。

步骤 4　遍历结束后，最大 AUC(Area Under the Curve)对应的子集 $\boldsymbol{P}$ 被认定为最佳特征范围子集。

4.2 特征降维

本文提出的隐式身份认证机制被期望基于超小规模的训练样本实现，而直接将前述操作处理后的小规模训练集用于模型拟合时，分类器的 AUC 指标存在大幅下降的现象。

由于传感器采集到的数据维度越高，其噪声越多，信息冗余比例就越大。如果出于提升性能的目的而直接对当前的小规模高维数据集进行数据扩增的操作，则分类器各项性能指标反而会出现一定程度的下降。为了减轻数据中的噪声对后续处理的影响，同时不过度损失原有高维数据中包含的信息，本文采用 PCA(Principal Component Analysis)算法对原有数据进行特征降维的操作。其实现过程如下：

步骤 1　令输入的用户行为特征向量 $\boldsymbol{F}'=\{p_1',p_2',p_3',\cdots,p_{21}',\cdots,p_{20n}'\}$，对其进行去中心化的操作。

步骤 2　计算协方差矩阵 $\frac{1}{n}\boldsymbol{F}'\boldsymbol{F}'^{\mathrm{T}}$，并利用特征值分解的方法求协方差矩阵的特征值与特征向量。

步骤 3　对特征值从大到小排序，选择特征值头部合适的 m 个，并将其所对应的 m 个特征向量作为行向量组成特征向量矩阵。最后，将原有数据转换到前述特征向量矩阵构建的样本空间中，得到经过特征降维处理的行为特征向量，其长度 $m\ll 20n$。

4.3 数据扩增

完成前述处理后，数据已经可以作为训练集导入分类器进行训练。但是，由于样本数较少，单类支持向量机在样本空间中创设出的可接受样本边界的边缘会出现偏移。SMOTE(Synthetic Minority Oversampling Technique)算法[12]能够对少数类样本进行上采样操作，其提出之初是为了解决两类样本的不平衡分布现象，目前鲜有 SMOTE 算法在隐式身份认证识别中扩展样本数据的应用，也鲜有使用其进行单类样本处理的工作，但考虑到 SMOTE 算法为类内运算，故对其活用以处理单类样本数据。本方案使用其改进型 SVM SMOTE 来增强扩展样本的正交性，具体实现过程如下：

步骤 1　令训练样本集 $\boldsymbol{S}=\{\boldsymbol{F}_1'\parallel\boldsymbol{F}_2'\parallel\boldsymbol{F}_3'\parallel\cdots\boldsymbol{F}_p'\}$，其中 p 为用户重复录入的次数。构造 $\boldsymbol{S}'=\{\boldsymbol{B}_1'\parallel\boldsymbol{B}_2'\parallel\boldsymbol{B}_3'\parallel\cdots\boldsymbol{B}_q'\}$，其中 $\boldsymbol{B}_n'=\{0,0,0,\cdots,0\}$ 为长度与 $\boldsymbol{F}'$ 相同的全零向量。设定补足参数 q 满足 $q\gg p$。

步骤 2　对 $\{\boldsymbol{S},\boldsymbol{S}'\}$ 应用 SVM SMOTE 算法，算法会将少数类的 $\boldsymbol{S}$ 样本集补足至其所含元素数量与 $\boldsymbol{S}'$ 相等，即获得训练样本集 $\boldsymbol{S}_{\mathrm{aug}}=\{\boldsymbol{F}'_{\mathrm{aug}\,1}\parallel\boldsymbol{F}'_{\mathrm{aug}\,2}\parallel\boldsymbol{F}'_{\mathrm{aug}\,3}\parallel\cdots\boldsymbol{F}'_{\mathrm{aug}\,p}\parallel\cdots\boldsymbol{F}'_{\mathrm{aug}\,q}\}$。其中 $\boldsymbol{F}'_{\mathrm{aug}\,i}$ 为经过上采样插值后得到的行为数据向量，$i\in[1,q]$。

步骤 3　去除 $\boldsymbol{S}'$ 后余下的 $\boldsymbol{S}_{\mathrm{aug}}$ 即为所求扩增后的训练样本集。

此时的训练样本集 $\boldsymbol{S}_{\mathrm{aug}}$ 中的元素数量较原有样本集 S 得到了较大程度的提升，且由插值得到的用户行为数据样本，即

样本空间中的支持向量之间保持了较好的正交性。由于单分类算法往往对数据的分布不进行假设，而对数据数量存在依赖性，因此此种插值的方案能够较为有效地提升单分类器的分类性能。

4.4 身份认证

本文所提机制基于单分类器，在训练阶段仅需要机主数据，从而将传统的分类问题转化为异常检测任务。即，在身份识别的过程中，将经过前述处理的训练样本集$\boldsymbol{S}_{\text{aug}}$用于分类器的训练，得到充分训练的分类器模型。在后续解锁时，设备利用其上搭载的各项传感器对用户的各项行为数据进行刻画，并生成该用户的行为数据样本。随后，设备使用分类器模型对样本进行匹配以得到用户身份。

为了确保选择的分类器最优，本文对常用的 3 种单分类器进行性能比较。使用单类支持向量机(One-class SVM)、隔离森林(Isolation Forests)分类器、局部异常因子(Local Outlier Factor)分类器对前述处理好的数据样本进行拟合，性能评估结果如表 1 所列。

由表 1 可得，基于前述的数据集，OCSVM 相比两种对比方案在准确率上体现出了 3%～6%的优势，在 FAR(False Acceptance Rate)上降低了约 30%，在 FRR(False Rejection Rate)上甚至可以体现出量级的差距，在综合衡量模型性能指标的 AUC 上同样也体现出了约 15%的优势。由此，选定 OCSVM 作为本方案的分类器算法。

在确定上述方案后，进一步地，使用网格搜索(Grid Search)方案对单类支持向量机的最佳参数进行搜索。

表 1 3 类单类分类器的性能比较

Table 1 Performance comparison among three one-class classifiers

Pattern	Classifier	*Accuracy*/%	*FAR*/%	*FRR*/%	*AUC*
Pattern L	Isolation Forests	96.97	41.57	2.72	0.7785
	Local Outlier Factor	96.60	47.69	3.05	0.7463
	One-Class SVM	99.34	17.87	0.53	0.9082
Pattern Z	Isolation Forests	90.96	37.92	8.79	0.7665
	Local Outlier Factor	93.91	44.58	5.77	0.7483
	One-Class SVM	96.56	16.39	3.33	0.9013
Pattern S	Isolation Forests	94.89	45.14	4.79	0.7504
	Local Outlier Factor	91.48	39.21	8.27	0.7625
	One-Class SVM	96.98	14.91	2.92	0.9110
Pattern T	Isolation Forests	97.56	41.39	2.14	0.7824
	Local Outlier Factor	97.17	50.74	2.45	0.7340
	One-Class SVM	99.30	13.70	0.60	0.9286

5 性能测试与分析

5.1 实验设置

实验的数据采集平台为 Nova 2s 手机，通过在该平台上安装数据采集软件进行相应的数据采集工作。在数据采集的环节，一共采集了 42 名参与者(22 名男性与 20 名女性)共 12600 条原始数据，参与者的平均年龄为 26.5 岁（标准差为 9.3）。

参与者被要求在给定设备上绘制图 2 所示的 4 种手势密码图形各 25 次。在某一位用户的行为数据被选为训练数据时，其中 16 次解锁动作产生的数据被用于训练，剩余数据被用于测试。

在测试的过程中，使用了三折交叉验证以避免出现过拟合的情况。此外，实验程序会依序将每一位用户的数据作为合法用户数据，从其余 41 名用户的数据中选择 17 名用户的数据作为非法用户测试数据，计算各项性能指标的平均值并将其作为最终结果，直至用户组遍历完毕。

在数量上，本文方案在每一组实验中的训练过程仅基于 16 条未标注的行为数据$\boldsymbol{S}_{\text{unlabeled}}=\{\boldsymbol{F}_1;\boldsymbol{F}_2;\cdots;\boldsymbol{F}_{16}\}$；作为对比，传统二分类器使用 735 条标注数据$\boldsymbol{S}_{\text{labeled}}=\{\boldsymbol{F}_1;\boldsymbol{F}_2;\cdots;\boldsymbol{F}_{735}\}$进行训练。

5.2 实验结果

5.2.1 性能对比

首先对本文特征处理方案中的操作使用准确率、*FAR*、*FRR*、*AUC* 4 种性能指标进行测试。对于特征排序、筛选的步骤，不同类别的手势密码数据在处理前后的特征维度变化如图 4 所示。以 S 型手势为例，使用单类支持向量机分类器分类时，特征排序、筛选前后的该类别数据分类性能的 4 项指标对比如图 5 所示。特征降维后，不同类别数据的维度变化如图 6 所示。在数据扩增的步骤中，同样以 S 型手势、单类支持向量机为例进行测试，该类别数据分类性能的 4 项指标如图 7 所示。

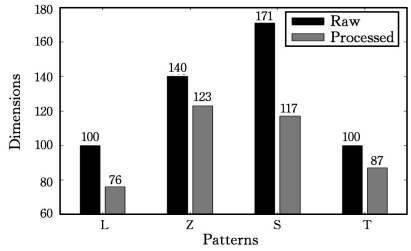

图 4 特征排序、筛选前后的各类别数据特征维度变化

Fig. 4 Dimensional changes of data of four patterns after feature sorting and filtering

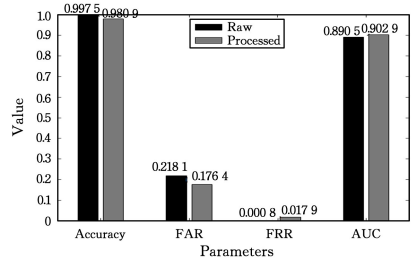

图 5 特征排序、筛选前后的 S 型手势数据分类性能比较

Fig. 5 Classification performance comparison about feature sorting and filtering on S pattern

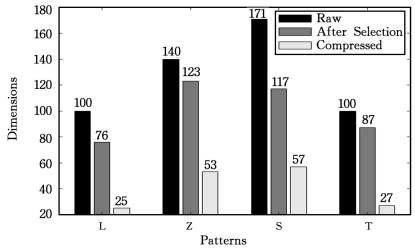

图 6 特征降维前后的各类别数据维度变化

Fig. 6 Dimensional changes of data of four patterns after dimension compression

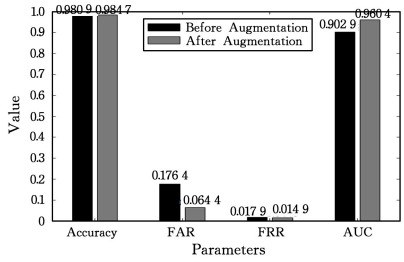

图 7 数据扩增后的 S 型手势数据分类性能比较

Fig. 7 Classification performance comparison about dimension compression and data augmentation on S pattern

由实验结果可知，完成特征排序与筛选后，各类手势密码的部分特征被直接抛弃，维度出现了小幅下降，*AUC* 值出现了一定的上浮。完成特征降维后，数据的高维特征被投影到低维空间上，维度出现了大幅度下降，从而减弱了噪声在后续步骤中的影响。通过上采样方案完成数据扩增后，*FAR* 指标大幅下降。在使用单类支持向量机的前提下，所使用的测试数据集中正样本远少于负样本，故 *FRR* 值变化不大。此时，作为模型性能的直观体现，*AUC* 值出现了 6%的增幅。

为了评估本文方案的性能，按照实验设置对本文方案与传统二分类器展开性能对比测试。

首先，对所使用的数据集测试了数种业界常用的二分类器，如逻辑回归(Logistic Regression)分类器、随机森林(Random Forests)分类器、KNN(K-Nearest Neighbor)分类器、梯度提升决策树(Gradient Boosting Decision Tree)分类器、多项式朴素贝叶斯(Multinomial Naive Bayes)分类器。

在上述二分类器中，在 4 种不同的图形密码下，KNN 分类器均表现出了一致的最佳性能。考虑到二分类并非本文的工作重点，本文不给出二分类器间的性能测试结果。依据上述结论，将使用 735 条标注行为数据 $\boldsymbol{S}_{\text{labeled}}$ 进行训练的 KNN 二分类器与使用 16 条未标注行为数据 $\boldsymbol{S}_{\text{unlabeled}}$ 进行训练的本文方案进行性能对比，测试结果如表 2 所列。

表 2 本文方案与传统 KNN 分类器的性能测试结果

Table 2 Performance comparison between proposed scheme and traditional KNN classifier

Pattern	Classifier	*Accuracy*/%	*FAR*/%	*FRR*/%	*AUC*
Pattern L	K Nearest Neighbor	99.51	12.61	0.19	0.936 1
	One-Class SVM	98.53	6.76	1.43	0.959 1
Pattern Z	K Nearest Neighbor	99.75	5.71	0.11	0.970 9
	One-Class SVM	97.50	10.88	2.43	0.933 3
Pattern S	K Nearest Neighbor	99.73	6.46	0.11	0.967 3
	One-Class SVM	97.70	4.54	2.28	0.965 9
Pattern T	K Nearest Neighbor	99.81	4.44	0.08	0.977 5
	One-Class SVM	98.18	5.65	1.79	0.962 9

由表 2 可见，对于 4 种手势图形密码，本文方案的准确率、*FAR*、*FRR*、*AUC* 4 项指标与使用大规模数据训练的 KNN 分类器接近。其中，对于覆盖点数较多但转折也较多的 S 型手势与 T 型手势，本文方案的 *AUC* 与对比方案的差距在 0.01 左右；对于覆盖点数较少的 L 型手势，本文方案的 *AUC* 甚至优于对比方案；对于覆盖点数较多但转折较少的 Z 型手势，本文方案的 *AUC* 与对比方案的差距为 0.04。整体而言，应用本文方案后，即使用于训练的初始样本数量较少，用户端的使用体验也不会发生明显变化。

由文献[13]可知，手势密码存在认证规律：直线部分的距离越大，识别性能越好；折线部分的拐点越多，识别性能越好。就本文选用的 4 种手势图形密码而言，L，S，T 型手势中转折所占比重较大，由转折点贡献的信息较多，因此在本文方案下分类性能较好；Z 型手势因为占用点数相对较多，但转折的次数仅有两次，由转折点贡献的信息成分不足，而直线段不具有太多的可用特征信息，因此仅依据 16 条行为数据的分类性能相对欠佳。

5.2.2 输入样本数对准确率的影响

在 5.2.1 节的性能测试中，本文方案仅选用 16 条未标注行为数据进行训练，即用户在录入阶段需要重复进行 16 次手势解锁的操作。进一步地，本文额外补充实验以确定用于训练的原始未标注数据条数对分类器性能表现的影响。

实验分别使用 8 组、10 组、12 组、14 组、16 组未标注行为数据作为训练原始样本，并按照前述的数据处理方案进行处理后，导入前述的隐式身份认证机制中进行性能评估。不同组数的输入样本对各项性能指标的影响如图 8—图 11 所示。

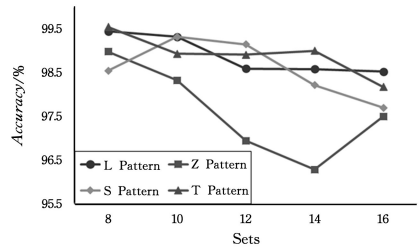

图 8 输入样本对准确率的影响

Fig. 8 Performance impact on accuracy caused by number of input samples

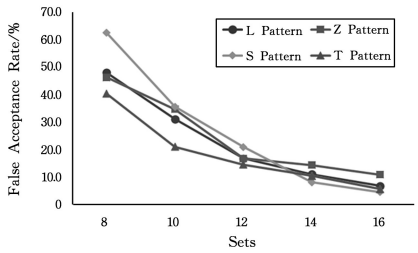

图 9 输入样本对 *FAR* 的影响

Fig. 9 Performance impact on *FAR* caused by number of input samples

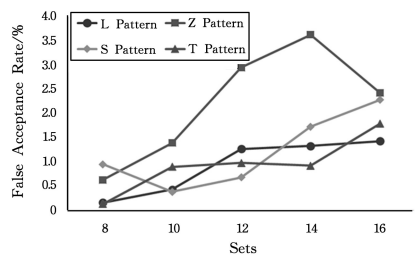

图 10 输入样本对 *FRR* 的影响

Fig. 10 Performance impact on *FRR* caused by number of input samples

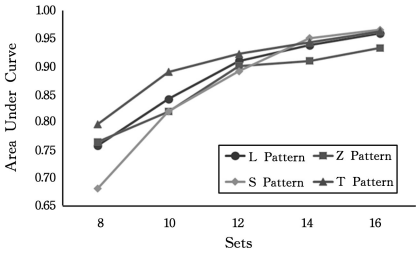

图 11 输入样本对 *AUC* 的影响

Fig. 11 Performance impact on *AUC* caused by number of input samples

由图 8—图 11 不难得出，*FAR*，*FRR*，*AUC* 均与原始样本数量体现出一定的相关性。即 *FAR*，*FRR* 两项指标随着原始样本数量的增长体现出下降的趋势，而 *AUC* 随着原始样本数量的增长体现出上升的趋势。特别地，可以注意到 Z 型手势在准确率与 *FRR* 上体现出了一定的波动，我们认为这种波动现象是由于该种手势覆盖点数较多而转折较少（仅为两次），所含有效信息不足[13]而由特定样本导致的。此外，由于测试数据集中正样本的数量远少于负样本的数量，由正样本使用数量带来的 *FRR* 值的小幅波动是可以忽略的。这里的重点应该主要放在体现模型有效性的综合指标 *AUC* 上。

总体来说，在所提的隐式身份认证机制录入阶段，通过增加用户重复录入的次数，可以在一定程度上提升分类器的性能指标，防止非设备拥有者访问敏感信息，从而提升安全性。

结束语 当下关于智能手机的隐式身份认证机制研究仍停留在早期阶段，大多数现存的相关研究仅关注模型的性能表现而非方案的实用性。为了解决相关工作的不足，本文提出的基于数据扩增的单分类认证方案对于小样本集下的分类性能具有改善作用。

实验结果表明，通过进行特征筛选、数据扩增的处理，即使用于训练的初始样本仅有单类且数量较少，相比较大规模数据训练的 KNN 分类器，所提方案在转折占比较大的手势上性能优异，在转折占比较小的手势上性能差距亦不明显，用户端的使用体验不会发生明显变化，充分表明所提方案有效可行。

参 考 文 献

[1] LEE M K. Security notions and advanced method for human shoulder-surfing resistant PIN-entry [J]. IEEE Transactions on Information Forensics and Security, 2014, 9(4): 695-708.

[2] SCHNEEGASS S, STEIMLE F, BULLING A, et al. SmudgeSafe: geometric image transformations for smudge-resistant user authentication [C] // Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp '14). 2014: 775-786.

[3] HOANG T, NGUYEN T C, LUONG C, et al. Adaptive cross-device gait recognition using a mobile accelerometer [J]. Journalof Information Processing Systems, 2013, 9(2): 333-348.

[4] SITOVAZ, SEDENKA J, YANG Q, et al. HMOG: new behavioral biometric features for continuous authentication of smartphone users [J]. IEEE Transactions on Information Forensics and Security, 2016, 11(5): 877-892.

[5] LIU L C, LI R G, YIN L H, et al. Research on multi-feature fusion impact authentication for intelligent mobile device[J]. Acta Electronica Sinica, 2016, 44(11): 2713-2719.

[6] KONGJ, GUO Y B, LIU C H, et al. Gait feature identification method based on motion sensor in smartphone[J]. Journal of Computer Applications, 2019, 39(6): 1747-1752.

[7] XIANG D D, CHEN H G, XIONG J J. Research on identity verification of mobile based on user behavior characteristics [J]. Journal of Shanghai Normal University (Natural Sciences), 2019, 48(2): 151-159.

[8] WANG R Z, TAO D. Context-aware implicit authentication of smartphone users based on multi-sensor behavior [J]. IEEE Access, 2019, 7: 119654-119667.

[9] YANG Y F, GUO B, WANG Z, et al. BehaveSense: Continuous authentication for security-sensitive mobile apps using behavioral biometrics [J]. Ad Hoc Networks, 2019, 2019(84): 9-18.

[10] SUN C, WANG Y, ZHENG J. Dissecting pattern unlock : The effect of pattern strength meter on pattern selection [J]. Journal of Information Security and Applications, 2014, 19: 4-5.

[11] SHI X P, WONG Y D, LI M Z, et al. A feature learning approach based on XGBoost for driving assessment and risk prediction [J]. Accident Analysis & Prevention, 2019, 129: 170-179.

[12] DONG M G, JIANG Z L, JING C. Multi-class imbalanced learning algorithm based on hellinger distance and smote algorithm [J]. Computer Science, 2020, 47(1): 102-109.

[13] LIU B Y. Smart phone identity authentication based on gesture recognition [D]. Beijing: Beijing Jiaotong University, 2018.

YAO Mu-yan, born in 1998. His main research interests include big data analysis, privacy and security.

TAO Dan, born in 1978, Ph.D, professor, Ph.D supervisor, is a senior member of China Computer Federation and a member of CCF TCIoT. Her main research interests include the IoT, mobile computing, wireless network, big data security and intelligent information processing.