·开发研究与工程应用· 文章编号：1000-3428(2021)04-0277-08 文献标志码：A 中图分类号：TP393

基于特征组分层与半监督学习的鼠标轨迹识别

康璐璐[1,3]，范兴容[2]，王茜竹[1,3]，杨晓雅[1,3]，明 蕊[1]

（1.重庆邮电大学 通信与信息工程学院，重庆 400065；2.重庆工商大学 计算机科学与信息工程学院，重庆 400067；
3.重庆邮电大学 电子信息与网络工程研究院，重庆 400065）

摘 要：传统时间序列分类方法存在鼠标轨迹特征挖掘不充分、数据不平衡与标记样本量少等问题，造成识别效果较差。结合特征组分层和半监督学习，提出一种鼠标轨迹识别方法。通过不同视角构建有层次的鼠标轨迹特征组，并借鉴半监督学习的思想，利用多个随机森林模型对未标记样本进行伪标记，且将抽取标签预测一致且置信度较高的部分样本加入到训练集中。基于基础特征组和辅助特征组，在扩充后的训练集上训练随机森林模型，以实现鼠标轨迹的人机识别。实验结果表明，该方法可有效识别鼠标轨迹，且精确率、召回率与调和均值分别达到97.83%、94.72%和96.56%。

关键词：鼠标轨迹识别；特征组分层；半监督学习；随机森林模型；不平衡数据

开放科学（资源服务）标志码（OSID）：

中文引用格式：康璐璐，范兴容，王茜竹，等.基于特征组分层与半监督学习的鼠标轨迹识别[J].计算机工程，2021，47(4)：277-284.

英文引用格式：KANG Lulu，FAN Xingrong，WANG Qianzhu，et al. Mouse trajectory recognition based on feature group hierarchy and semi-supervised learning[J]. Computer Engineering，2021，47(4)：277-284.

Mouse Trajectory Recognition Based on Feature Group Hierarchy and Semi-Supervised Learning

KANG Lulu[1,3]，FAN Xingrong[2]，WANG Qianzhu[1,3]，YANG Xiaoya[1,3]，MING Rui[1]

（1.School of Communication and Information Engineering，Chongqing University of Posts and Telecommunications，Chongqing 400065，China；2.School of Computer Science and Information Engineering，Chongqing Technology and Business University，Chongqing 400067，China；3.Electronic Information and Networking Research Institute，Chongqing University of Posts and Telecommunications，Chongqing 400065，China）

【Abstract】Traditional time series classification methods have problems such as insufficient mining of mouse trajectory features，unbalanced data，and few labeled samples，resulting in poor recognition results. Combining feature group hierarchy and semi-supervised learning，this paper proposes a mouse track recognition method. In this method，hierarchical mouse trajectory feature groups are constructed from different perspectives.Then based on the idea of semi-supervised learning，multiple random forest models are used to pseudo-label unlabeled samples，and some samples with consistent label predictions and high confidence are added to the training set.Based on the basic feature set and auxiliary feature set，the random forest model is trained on the expanded training set to realize the human-machine recognition of the mouse trajectory. The experimental results show that this method can effectively identify the mouse track，and its precision，recall rate and harmonic mean values reach 97.83%，94.72% and 96.56%，respectively.

【Key words】mouse trajectory recognition；feature group hierarchy；semi-supervised learning；random forest model；unbalanced data

DOI：10. 19678/j. issn. 1000-3428. 0057442

基金项目：重庆市自然科学基金（cstc2018jcyjAX0587）；重庆市科技重大主题专项重点示范项目（cstc2018jszx-cyztzxX0035）；中国移动科研基金项目（MCM20170203）。

作者简介：康璐璐（1994—），女，硕士研究生，主研方向为机器学习、数据挖掘；范兴容，讲师、博士；王茜竹，高级工程师、博士；杨晓雅，硕士研究生；明 蕊，本科生。

收稿日期：2020-02-20 **修回日期：**2020-03-27 **E-mail：**1799393698@qq.com

0 概述

随着互联网行为式验证码技术的快速发展,以拖动滑块为代表的鼠标轨迹识别因其传输数据小、暴力破解难度大等特点,已经广泛运用于多种人机验证产品中。然而,攻击者仍可通过黑产工具产生的类人轨迹批量操作来躲避检测,并在对抗过程中不断升级其伪造数据以持续躲过同样升级的检测技术。因此,研究一种新型的鼠标轨迹识别技术来增强人机攻防识别性能,在保障网站抵御黑客攻击、用户恶意大规模访问与机器批量在线投票等方面具有重要的现实意义。

鼠标轨迹是用户在使用鼠标拖动滑块的过程中,经过采样获得水平方向、垂直方向和时间3个维度的轨迹点集。鼠标轨迹时间序列数据相较传统时间序列具有以下6个特点:多变量,即鼠标轨迹包括水平方向x轴、垂直方向y轴和时间t轴3个维度;不规则采样,即由于网络延时等原因使得每个采样点之间的时长不同;长度不等,即由于鼠标轨迹采样间隔不确定,导致每一条轨迹的长度不等;变量之间存在关联性,即x、y、t3个维度在时间和空间上存在关联性;数据不平衡,即人类轨迹样本数远多于机器轨迹样本数;标记样本少,即考虑到标记数据获取困难、标记代价高等问题,导致训练样本数量少。因此,鼠标轨迹识别可以看作一种特殊的时间序列二分类问题,同时也是一个典型的人机识别问题。

传统时间序列分类方法[1-3]不能直接用于鼠标轨迹的识别,如鼠标轨迹时间序列的长度不等,将会导致其不能直接作为传统时间序列分类方法的输入。虽然动态时间规整(Dynamic Time Warping,DTW)方法[4]能够处理长度不同的时间序列分类,但是由于鼠标轨迹序列的采样时间不等,因此不能进行准确的序列相似性判断。为此,本文对鼠标轨迹数据进行充分挖掘,从不同视角提取基础特征组和辅助特征组,提出一种基于特征组分层和半监督学习的鼠标轨迹识别方法。利用半监督随机森林算法对标记样本进行扩充,采用随机抽样改善数据不平衡问题,并通过将半监督随机森林算法与特征组分层策略相结合来提升鼠标轨迹的识别性能。

1 相关工作

目前,针对不规则采样与长度不等的时间序列分类问题主要有两类解决方法。

第一类是基于模型的方法,如文献[5]通过高斯过程(Gaussian Process,GP)后验重新表示每个时间序列,然后在GP后验空间上定义核函数,并应用基于核函数的方法进行时间序列分类。文献[6]对肾小球滤过率时间序列进行自动分类,先采用高斯回归过程填补特殊时间序列的缺失值,并转换为长度相等的时间序列,再使用KNN/SVM算法对其分类。文献[7]从单个时间序列中推导出时间连续的动力系统模型,并用推导出的模型表示时间序列,使用分类器对模型上的后验分布进行分类。上述研究在实验中均取得满意的效果,但这些方法都是基于二维时间序列,且没有提出针对数据不平衡以及标记样本量较少的处理方法,因此不适用于本文鼠标轨迹数据具有的多变量、变量之间存在关联性、数据不平衡与标记样本量少等实际情况。

第二类是基于特征的方法,该方法用一组特征来表征时间序列信息,从而解决时间序列不规则问题。如文献[8]提出一种基于特征的时间序列分类方法,该方法从时间序列中提取数千个可解释的特征,并使用贪婪前向特征选择方法选择出信息量最大的特征。文献[9]将时间序列预测与不同的特征选择相结合,实现一个更为简单的建模过程。针对鼠标轨迹识别的实际问题,目前国内外的研究都较少,且多数采用基于特征的分类方法。如文献[10]对提取的鼠标行为特征进行分析,证明了利用行为特征进行身份认证的可行性。文献[11-13]基于有监督学习构建出鼠标轨迹识别方法,利用轨迹信息分别构建特征工程,并采用梯度提升模型与朴素贝叶斯模型等模型进行人机识别,且取得较好的识别结果。但这些方法存在特征挖掘不充分且没有考虑类别不平衡等问题,使得模型的泛化性能较弱。文献[14]提出一种融合并行投票决策树和半监督学习的鼠标轨迹识别方法,该方法提取出105个鼠标轨迹特征,采用基于半监督的策略与并行决策投票树的思想进行识别。然而,该方法仍然不能解决数据不平衡问题,且特征工程过于冗余,造成识别效果有限。

针对上述方法存在的特征挖掘不充分且鼠标轨迹数据存在标记样本少、数据不平衡等问题,本文提出一种基于特征组分层和半监督学习的鼠标轨迹识别方法。该方法通过构建有层次的鼠标轨迹特征组,描述用户滑动鼠标轨迹的移动规律,增加鼠标轨迹识别置信度。借鉴半监督学习的思想,利用多个随机森林模型对未标记样本进行伪标记,抽取标签预测一致且高置信度的部分样本加入到训练集合中,再基于基础特征组和辅助特征组,在扩充的训练样本集上重新训练随机森林算法,实现鼠标轨迹的人机识别。

2　背景工作

2.1　数据样本

本文的实验数据来源于某人机验证产品经过脱敏后的鼠标轨迹数据[15]，其主要字段为唯一编号id、鼠标轨迹水平坐标x、鼠标轨迹垂直坐标y、鼠标轨迹采样时间t、轨迹目标点水平坐标x_a和轨迹目标点垂直坐标y_a。鼠标轨迹的数据字段说明如表1所示。

表1　鼠标轨迹数据字段说明

Table 1　Field description of mouse trajectory data

字段	字段说明
id	区分不同鼠标轨迹的识别码
x	鼠标移动过程中采样获得的水平坐标
y	鼠标移动过程中采样获得的垂直坐标
t	鼠标轨迹的采样时间
(x_a, y_a)	鼠标轨迹目标点的水平坐标和垂直坐标

2.2　特征提取

本文分别从描述人类轨迹特性和强化人机轨迹差异性的角度来提取基础特征组和辅助特征组。

2.2.1　基于人类轨迹特性的基础特征组

基础特征组是从基于人类鼠标轨迹特性的角度而构建的，这是因为人类轨迹特征具有较好的稳定性，且主要体现在以下3个方面：

1）拟合过程：人们拖动滑块接近目标位置时存在拟合现象，会缓慢将滑块放到目标位置（拖拽速度逐渐变小），而机器轨迹速度则不会有明显的变化趋势。不同鼠标轨迹的水平移动速度如图1所示。

2）无规律性：人类轨迹的移动偏移量是不停变化的，而机器轨迹受黑客程序控制，其偏移量一般有规律可寻，具体如图2所示。

3）回退现象：人们在拖动滑块接近目标位置时，由于惯性会出现拖离目标点又重新拖动回来的现象，而机器轨迹到达目标点后则会立即停止，不存在回退现象，具体如图3所示。

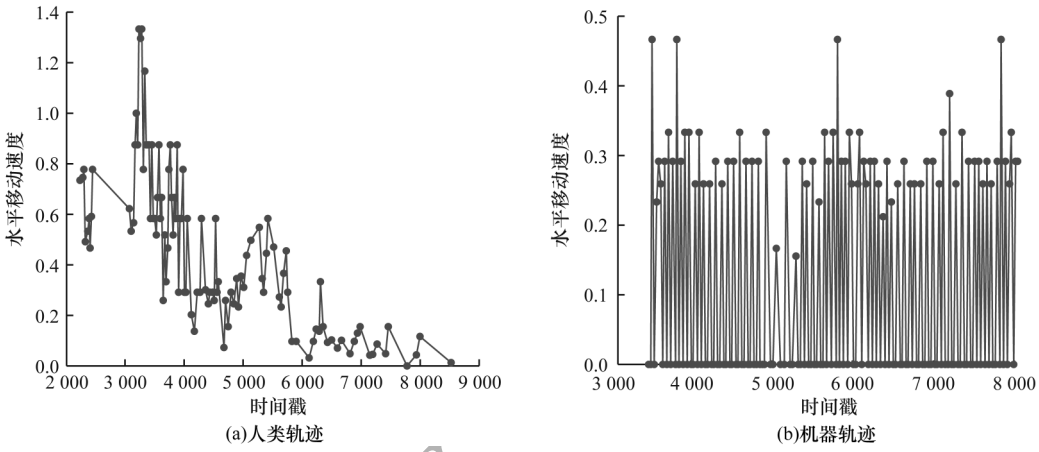

图1　人类轨迹与机器轨迹的水平移动速度

Fig.1　Horizontal moving speed of human trajectory and machine trajectory

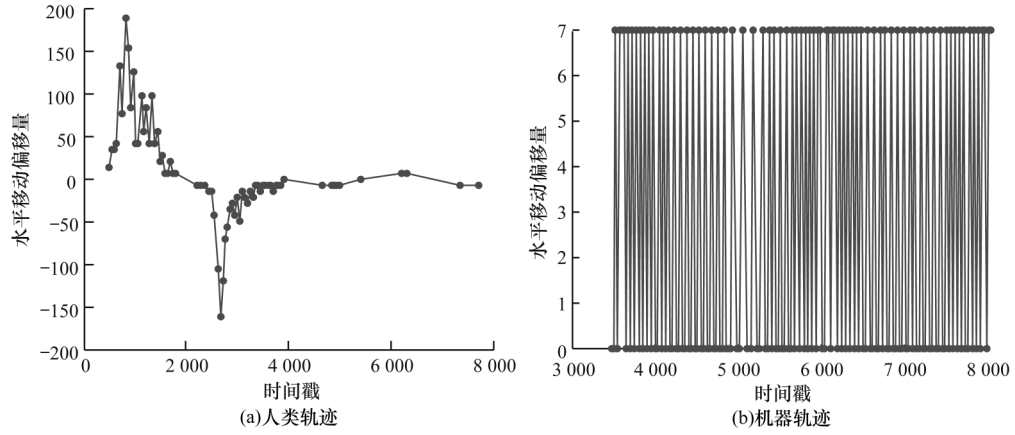

图2　人类轨迹与机器轨迹的水平移动偏移量

Fig.2　Horizontal movement offset of human trajectory and machine trajectory

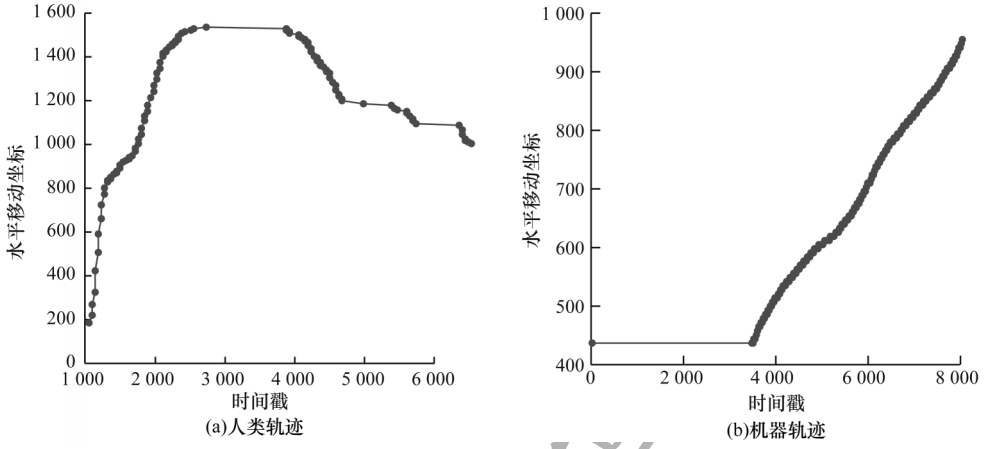

图3 人类轨迹与机器轨迹的回退现象

Fig.3 Regression phenomenon of human trajectory and machine trajectory

基于上述规律,本文主要提取以下基础特征:

1)水平坐标最大值与目标值之差x_{ovs}和水平坐标最大值与最小值之差x_{differ}。假设$\boldsymbol{X}=[x_1,x_2,\cdots,x_n]$为轨迹水平坐标($n$为鼠标轨迹采样点数),则有:

$$x_{ovs}=\max(x_1,x_2,\cdots,x_n)-x_a \tag{1}$$

$$x_{differ}=\max(x_1,x_2,\cdots,x_n)-\min(x_1,x_2,\cdots,x_n) \tag{2}$$

其中,x_a为轨迹目标点的水平坐标,$\max(x)$函数和$\min(x)$函数用来返回数组的最大值与最小值。

2)水平坐标偏移量最小值$\mathrm{d}x_{min}$和水平坐标偏移量标准差$\mathrm{d}x_{std}$。假设$\mathrm{d}\boldsymbol{X}=[\mathrm{d}x_1,\mathrm{d}x_2,\cdots,\mathrm{d}x_{n-1}]$为轨迹水平坐标的偏移量,则有:

$$\mathrm{d}\bar{x}=\frac{\mathrm{d}x_1+\mathrm{d}x_2+\cdots+\mathrm{d}x_{n-1}}{n-1} \tag{3}$$

$$\mathrm{d}x_{min}=\min(\mathrm{d}x_1,\mathrm{d}x_2,\cdots,\mathrm{d}x_{n-1}) \tag{4}$$

$$\mathrm{d}x_{std}=\sqrt{\frac{(\mathrm{d}x_1-\mathrm{d}\bar{x})^2+(\mathrm{d}x_2-\mathrm{d}\bar{x})^2+\cdots+(\mathrm{d}x_{n-1}-\mathrm{d}\bar{x})^2}{n-1}} \tag{5}$$

3)回退轨迹(拖离目标点后重新拖动回来产生的轨迹,无回退轨迹则取轨迹后10个点)水平坐标偏移量最小值$\mathrm{d}x'_{min}$和回退轨迹水平坐标偏移量中程数$\mathrm{d}x'_{mid}$。假设$\boldsymbol{X}_e=[x'_1,x'_2,\cdots,x'_m]$为回退轨迹水平坐标($m$为回退轨迹采样点数),$\mathrm{d}\boldsymbol{X}_e=[\mathrm{d}x'_1,\mathrm{d}x'_2,\cdots,\mathrm{d}x'_{m-1}]$为其偏移量,则有:

$$\mathrm{d}x'_{min}=\min(\mathrm{d}x'_1,\mathrm{d}x'_2,\cdots,\mathrm{d}x'_{m-1}) \tag{6}$$

$$\mathrm{d}x'_{mid}=\frac{\min(\mathrm{d}x'_1,\mathrm{d}x'_2,\cdots,\mathrm{d}x'_{m1})+\max(\mathrm{d}x'_1,\mathrm{d}x'_2,\cdots,\mathrm{d}x'_{m1})}{2} \tag{7}$$

4)回退轨迹速度最大值v'_{max}和回退轨迹速度末尾值v'_{end}。假设$\boldsymbol{V}_e=[v'_1,v'_2,\cdots,v'_{m-1}]$为回退轨迹移动速度,则有:

$$v'_{max}=\max(v'_1,v'_2,\cdots,v'_{m-1}) \tag{8}$$

$$v'_{end}=\mathrm{end}(v'_1,v'_2,\cdots,v'_{m-1}) \tag{9}$$

其中,$\mathrm{end}(v)$函数用来返回数组末尾值。

5)回退轨迹速度偏移量最大值$\mathrm{d}v'_{max}$和回退轨迹点个数x'_{num}。假设$\mathrm{d}\boldsymbol{V}_e=[\mathrm{d}v'_1,\mathrm{d}v'_2,\cdots,\mathrm{d}v'_{m-2}]$为回退轨迹移动速度的偏移量,则有:

$$\mathrm{d}v'_{max}=\max(\mathrm{d}v'_1,\mathrm{d}v'_2,\cdots,\mathrm{d}v'_{m-2}) \tag{10}$$

$$x'_{num}=m \tag{11}$$

2.2.2 强化人机轨迹差异性的辅助特征组

辅助特征组是从基于强化人机轨迹差异的角度而构建的,主要提取不明显的人机差异性,但在数据规模较大时依然不能忽略的特征(如y维度和t维度的特征)用于辅助判断,增加轨迹识别置信度。

1)垂直坐标最小值y_{min}和垂直坐标改变次数y_{chg}。假设$\boldsymbol{Y}=[y_1,y_2,\cdots,y_n]$为轨迹垂直坐标,则有:

$$y_{min}=\min(y_1,y_2,\cdots,y_n) \tag{12}$$

$$y_{chg}=\mathrm{cnt}(y_1,y_2,\cdots,y_n) \tag{13}$$

其中,$\mathrm{cnt}(y)$函数用来返回数组中数据不重复的个数。

2)垂直坐标偏移量初始值$\mathrm{d}y_{init}$。假设$\mathrm{d}\boldsymbol{Y}=[\mathrm{d}y_1,\mathrm{d}y_2,\cdots,\mathrm{d}y_{n-1}]$为垂直坐标的偏移量,则有:

$$\mathrm{d}y_{init}=\mathrm{init}(\mathrm{d}y_1,\mathrm{d}y_2,\cdots,\mathrm{d}y_{n-1}) \tag{14}$$

其中,$\mathrm{init}(\mathrm{d}y)$函数用来返回数组初始值。

3)采样时间初始值t_{init},采样时间中位数t_{med}和鼠标第一次移动到目标点所需时间t_{aim}。假设$\boldsymbol{T}=[t_1,t_2,\cdots,t_n]$为轨迹采样时间,则有:

$$t_{init}=\mathrm{init}(t_1,t_2,\cdots,t_n) \tag{15}$$

$$t_{med}=\mathrm{median}(t_1,t_2,\cdots,t_n) \tag{16}$$

$$t_{aim}=t_a-t_1 \tag{17}$$

其中,t_a为鼠标第一次移动到目标点的时间,$\mathrm{median}(t)$函数用来返回数组中位数。

4)采样时间偏移量初始值$\mathrm{d}t_{init}$。假设$\mathrm{d}\boldsymbol{T}=[\mathrm{d}t_1,\mathrm{d}t_2,\cdots,\mathrm{d}t_{n-1}]$为时间的偏移量,则有:

$$\mathrm{d}t_{init}=\mathrm{init}(\mathrm{d}t_1,\mathrm{d}t_2,\cdots,\mathrm{d}t_{n1}) \tag{18}$$

3 本文识别方法

本文提出一种基于特征组分层和半监督学习的鼠标轨迹识别方法。假设$v=v_1\times v_2$为鼠标轨迹特征空间,其中v_1和v_2分别表示人类轨迹特性和人机轨

迹差异性2个视角。本文所提方法的实现框图如图4所示。

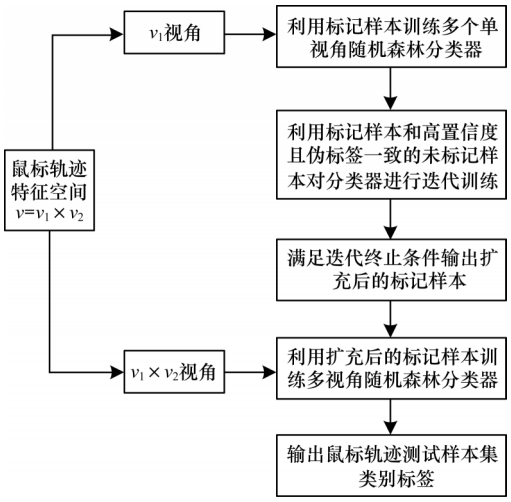

图4　本文所提方法的实现框图

Fig.4　Implementation block diagram of the proposed method

3.1　单视角随机森林分类器的训练过程

随机森林算法[16]具有训练速度快、模型泛化能力强以及可平衡误差等优势，因此本文采用随机森林作为基础分类算法。从v_1视角提取特征构成标记样本集L，将标记样本以自助(bootstrap)采样的方式抽取N份，单独训练N个单视角随机森林分类器$\{R_1,R_2,\cdots,R_N\}$，并使用这N个分类器对未标记样本U进行预测，其中任意一个分类器R_i对未标记样本的预测置信度和伪标签计算过程如下所示：

假设R_i中的单棵决策树$f_{\mathrm{id}}(x_u)=f(x_u,\theta_{\mathrm{id}})$，$\theta_{\mathrm{id}}$表示第$d$棵决策树构建过程中的随机因素(如特征的随机选择)，单视角随机森林分类器$R_i=\{f_{i1},f_{i2},\cdots,f_{iD}\}$，$D$为决策树个数，则分类器$R_i$将样本$x_u$预测为$k$类的概率为：

$$p_i(k|x_u)=\frac{1}{D}\sum_{d=1}^{D}p_{i,d}(k|x_u),k=0,1 \tag{19}$$

其中，$p_{i,d}(k|x_u)$为第d棵决策树中叶节点的类别预测概率。

样本x_u的预测置信度定义为：

$$\mathrm{Con}_i(x_u)=\max_{k\in C}p_i(k|x_u) \tag{20}$$

其中，$C\in\{0,1\}$表示样本类别集合。

样本x_u的伪标签为：

$$pl_i(x_u)=\operatorname*{argmax}_{k\in C}p_i(k|x_u) \tag{21}$$

3.2　单视角随机森林分类器的迭代过程

在单视角随机森林分类器的迭代过程中引入了半监督学习的思想[17-19]，并将部分未标记样本加入标记样本中，利用新的样本重新进行分类器的训练。

假设$\{R_1(x_u),R_2(x_u),\cdots,R_N(x_u)\}$为未标记样本$x_u$在$N$个分类器上的表示形式，则样本在$N$个分类器上的伪标签和预测为该标签的置信度分别为：

$$R_n(x_u)\rightarrow[pl_n(x_u),\mathrm{Con}_n(x_u)],n=1,2,\cdots,N \tag{22}$$

由于对样本的伪标记正确与否对后续多视角随机森林分类器的识别性能起至关重要的作用，因此本文从伪标签和置信度2个方面对未标记样本进行选择，即标记条件为：

1)样本在N个分类器中的伪标签一致，则$pl_1(x_u)=pl_2(x_u)=\cdots=pl_N(x_u)$。

2)样本在N个分类器中预测置信度大于阈值参数θ的个数至少有ε个，即$\mathrm{countif}(\mathrm{Con}_n(x_u)\geqslant\theta(n=1,2,\cdots,N))\geqslant\varepsilon$，其中countif为计数函数，若满足括号内条件则再加1。

由于数据样本类别的不平衡性(人类轨迹远多于机器轨迹)，如果将所有满足标记条件的未标记样本全部添加，可能会导致模型在第$\omega+1$轮更新后劣于第ω轮的模型。为了达到类别平衡的目的，按照标记样本类别之间的样本比率对未标记样本进行随机抽样，通过逐步缩小多数类别使得数据趋于平衡，具体抽样过程如下：

假设第ω轮迭代时标记样本集中人类样本和机器样本的比率为β，满足标记条件的未标记样本中人类样本B_1和机器样本B_2的数量分别为b_1和b_2，则有：

$$b_1'=\begin{cases}b_1, s_1<\dfrac{1}{\beta}\times b_2\\ \dfrac{1}{\beta}\times b_2, s_1\geqslant\dfrac{1}{\beta}\times b_2\end{cases} \tag{23}$$

$$b_2'=b_2 \tag{24}$$

$$B_i'=\mathrm{subsample}(b_i',B_i),i=1,2 \tag{25}$$

其中，b_1'和b_2'分别表示人类样本和机器样本的抽样个数，S_1'和S_2'表示最终要扩充进标记样本的人类样本集和机器样本集，$\mathrm{subsample}(b,B)$函数表示在B集中随机抽取b个样本。

将上述未标记样本及其伪标签加入到标记样本中，并从未标记样本中剔除。当进行分类器的迭代训练时直至满足迭代终止条件时停止。通过分类器的多次训练，逐步实现了标记样本的扩充，并改善了数据不平衡与标记样本量不足的问题。

3.3　多视角随机森林分类器的鼠标轨迹识别

经过多次迭代训练后，鼠标轨迹的标记样本集得到扩充，数据类别不平衡问题也得到改善，此时从$v_1\times v_2$视角提取特征构成多视角随机森林分类器训练样本$L'=\{(\boldsymbol{x}_i,\boldsymbol{y}_i,c_i)\}(i=1,2,\cdots,|L'|)$，其中$\boldsymbol{x}_i$和$\boldsymbol{y}_i$分别表示从视角$v_1$和视角$v_2$中提取的基础特征组和辅助特征组，$c_i$为类别标签，$|L'|$为标记样本和添加进的未标记样本的数量总和。可以看到此时样本的数量相比之前得到扩充，因此在特征中加入v_2视角，并利用此样本集训练多视角随机森林分类器，最终实现鼠标轨迹的人机识别，且识别算法如算法1所示。

算法1 基于特征组分层和半监督学习的鼠标轨迹识别方法

输入 基于视角v_1构建的初始标记样本集L,初始未标记样本集U,单视角随机森林分类器个数N,置信度阈值参数θ,置信度满足阈值条件的个数ε,本文$\varepsilon=\lceil N/2\rceil$

输出 鼠标轨迹测试样本集T预测标签

1.e′ ← 0.5//初始分类错误率
2.repeat
3.B_1←∅, B_2←∅
4.for i=1; i ⩽ N; i + + do
5.L_i← BootstrapSample(L)//自助法采样
6.R_i← $RF_i(L_i)$//RF为随机森林算法
7.end for
8.e←MeasureError(R_1 & R_2 &⋯& R_N)
9.if e < e′then
10.e′ ← e
11.for every $x_u \in U$ do
12.if $pl_1(x_u)=pl_2(x_u)=\cdots=pl_N(x_u)$then
13.if countif($Con_n(x_u)\geqslant\theta)\geqslant\varepsilon$ then
14.if $pl_N(x_u)$=1 then
15.$B_1 \leftarrow B_1 \cup (x_u, pl_N(x_u))$
16.else
17.$B_2 \leftarrow B_2 \cup (x_u, pl_N(x_u))$
18.end if
19.end if
20.end if
21.end for
22.end if
23.B_1', B_2' ← PseSample(B_1, B_2)
24.L ← L ∪ B_1' ∪ B_2', U ← U − B_1' − B_2'
25.until neither L nor U changes
26.在L中添加辅助特征组,构成L′
27.R → RF(L′)

3.4 使用R对测试样本集T进行预测

算法1中的PseSample为未标记样本中满足标记条件的样本抽样函数,MeasureError函数的作用是估计R_1 & R_2 &⋯& R_N组合的分类错误率。本文使用初始标记样本来计算分类错误率,假设该样本集中假设满足标记条件的样本有z个,其中有z'个的分类是正确的,则分类错误率$e=(z-z')/z$。如果第$\omega+1$次迭代训练后模型的分类错误率小于第ω次,则继续迭代,从而保证模型向性能提升的方向更新。

4 实验设置和评价准则

4.1 实验设置

1)数据设置。本文数据来源于某人机验证产品经过脱敏后的鼠标轨迹数据,经过数据筛选后,共有103 000条,其中人类轨迹82 600条,机器轨迹20 400条,数据比约为4∶1,可以看出数据类别存在不平衡现象。将其中3 000条数据样本用于模型训练,其中标记样本比例选用20%,未标记样本的比例选用80%,100 000条用于模型测试。

2)特征选择。实验共选取17个特征,包含描述人类轨迹特性的基础特征组10个,用f_1~f_{10}表示,强化人机轨迹差异的辅助特征组有7个,用f_{11}~f_{17}表示。

3)仿真平台。本文采用MATLAB2017a软件,在Windows7 64位操作系统Intel i5处理器的惠普电脑上进行实验测试。

4.2 评价准则

本文采用精确率P、召回率R与调和均值F_α作为模型的评价指标,计算方法如下所示:

$$P=\frac{TP}{TP+FP}\times 100\% \tag{26}$$

$$R=\frac{TP}{TP+FN}\times 100\% \tag{27}$$

其中,TP为被正确识别为机器轨迹的样本数,FP为被错误识别为机器轨迹的样本数,FN为被错误识别为人类轨迹的样本数。

为保障网站抵御黑客攻击等网络安全问题,通常要求尽可能地识别出机器轨迹(即偏重召回率),以避免漏识别机器轨迹导致不可挽回的损失,但又不能使得用户验证体验太差(保证较高的精确率)。因此,在衡量人机轨迹识别性能时引入了调和均值F_α,其计算方法为:

$$F_\alpha=\frac{(1+\alpha^2)PR}{\alpha^2P+R}\times 100\%=\frac{5PR}{2P+3R}\times 100\% \tag{28}$$

其中,$\alpha<1$表示合适的偏重召回率。

5 结果分析与讨论

5.1 不同特征的重要性分析

为评估不同特征对识别效果的影响,本文采用随机森林自带的评估特征重要性方法计算本文所提取特征f_1~f_{17}的重要性得分,结果如图5所示。

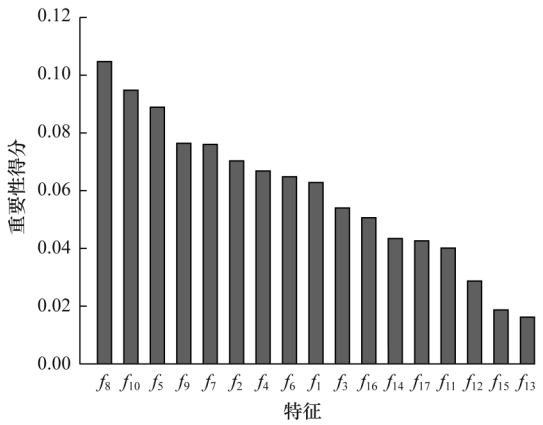

图5 本文提取特征的重要性得分

Fig.5 Importance score of features extracted in this paper

从图5可以看出,基础特征组f_1~f_{10}相较于辅助特征组f_{11}~f_{17},其总体重要性得分更高,这是因为基础特征组是基于人类鼠标轨迹特性的角度去构建

的，能够更好地区分人机轨迹，而辅助特征组主要用于辅助判断，增加轨迹识别置信度。

5.2　参数N和θ对识别性能的影响

文献[20]给出了随机森林分类算法中决策树各节点选取的特征个数m_{try}满足$m_{try}=\sqrt{d}$时，算法的性能最佳。其中，d为数据集的特征个数。在本文方法中，单视角随机森林分类器的参数$m_{try1}=3$、决策树个数$n_{tree1}=50$，多视角随机森林分类器的参数$m_{try2}=4$、决策树个数$n_{tree2}=100$。由于参数N和θ是影响本文算法性能的重要参数，因此通过对这2个参数设置不同的取值进行实验来找到参数的最优值。

图6给出了参数N和θ在不同取值下鼠标轨迹的识别结果。从图6可以看出，当$N=3,\theta=0.8$时，本文方法在召回率、精确率和调和均值上都具有较好的性能。如果参数过大，则未标记样本得不到充分利用，并且无法解决数据不平衡、标记样本不足的问题。如果参数过小，则会导致训练样本中引入过多的噪声数据，极大影响最终的识别效果。

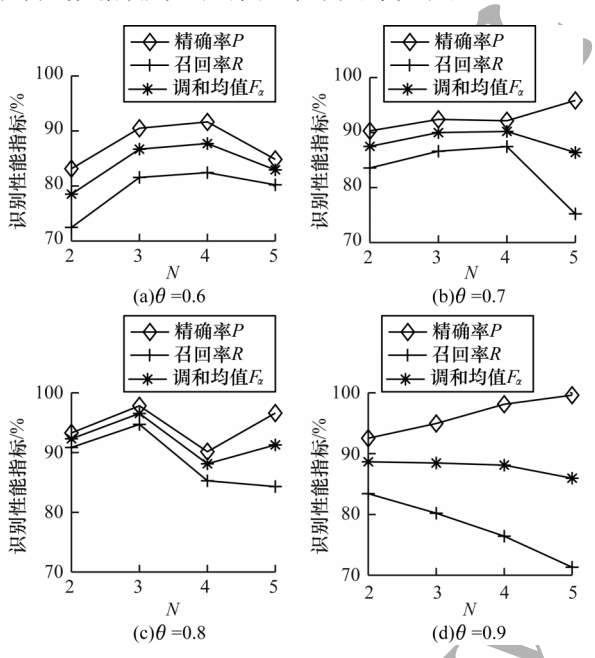

图6　参数N和θ对识别性能的影响

Fig.6　Effect of parameters N and θ on recognition performance

5.3　方法性能分析

5.3.1　特征组分层有效性分析

为了验证本文提出的特征组分层对鼠标轨迹识别性能的影响，实验分别比较了引入特征组分层和不引入特征组分层这两种方式下的识别性能，结果如表2所示。可以看出，引入特征组分层时模型的精确率、召回率和调和均值较不引入时分别提高了6.4、11.27和8.5个百分点，这说明在半监督学习的基础上引入特征组分层在鼠标轨迹识别中能够有效提高模型的识别性能。

表2　特征组分层有效性分析结果

Table 2　Analysis results of hierarchical effectiveness of feature group　%

方法	精确率P	召回率R	调和均值F_α
不引入特征组分层	91.43	83.45	88.06
引入特征组分层	97.83	94.72	96.56

5.3.2　方法运行时间分析

表3给出了参数N在不同取值下本文方法的训练耗时和测试耗时结果。可以看出，随着分类器个数N的增加，测试耗时相差不大，但训练耗时逐渐增加。因此，本文方法需要选取合适的N值，在保证分类性能的同时降低运行时间。

表3　分类器个数对本文方法运行时间的影响

Table 3　Effect of the number of classifiers on the running time of the proposed method　s

分类器个数	训练耗时	测试耗时
2	33.51	17.32
3	49.85	17.91
4	65.89	18.54
5	80.68	19.45

5.3.3　方法对比分析

实验将本文所提方法与基于朴素贝叶斯的鼠标轨迹识别方法[11]和基于梯度提升决策树的鼠标轨迹识别方法[12]在同一数据集上进行性能对比，结果如表4所示。其中，文献[11]提取8个特征，并使用加权朴素贝叶斯模型实现人机识别，文献[12]提取6个轨迹特征，使用梯度提升决策树模型对鼠标轨迹进行识别。由于这2种方法都采用监督学习，不涉及未标记样本的使用，因此在训练模型时只采用训练集中的标记样本。

表4　3种鼠标轨迹识别方法的性能对比

Table 4　Performance comparison of three mouse trajectory recognition methods　%

方法	精确率P	召回率R	调和均值F_α
文献[11]方法	89.51	82.35	86.50
文献[12]方法	91.73	78.13	85.76
本文方法	97.83	94.72	96.56

从表4可以看出：本文方法的精确率、召回率和调和均值较文献[11]分别提高8.32、12.37和10.06个百分点，较文献[12]分别提高6.1、16.59和10.8个百分点。造成上述结果主要有以下2个原因：1）文献[11-12]没有对轨迹特征进行充分挖掘，使得轨迹特征不能完整地刻画人机轨迹；2）本文方法采用半监督学习策略，充分利用了大量未标记样本提升模型性能，且通过随机抽样改善数据不平衡现象，而文献[11-12]采用有监督方法，未考虑鼠标轨迹数据不平衡和标记样本量少的实际情况，造成最终识别性能较低。

6 结束语

本文提出一种基于特征组分层和半监督学习的鼠标轨迹识别方法，该方法从特征和数据2个层面对鼠标轨迹识别方法进行改进。在特征层面，根据不同视角特征在不同阶段所起的作用构建出有层次的特征组并分层添加至模型中，避免在训练样本数量过少的情况下盲目添加特征而造成模型过拟合问题。在数据层面，利用半监督学习方法扩充训练样本，解决数据类别不平衡、标记样本量不足的问题，并通过将两者融合实现提升鼠标轨迹识别效果的目的。实验结果表明，本文方法能够有效提升鼠标轨迹识别任务的性能。针对未来黑客攻击呈现方式多元化的问题，下一步将采用深度学习动态提取特征，以应对黑客攻击方式的转变。

参考文献

[1] ABANDA A, MORI U, LOZANO J A. A review on distance based time series classification[J]. Data Mining and Knowledge Discovery, 2019, 33(2): 378-412.

[2] LINES J, DAVIS L M, HILLS J, et al. A shapelet transform for time series classification[C]//Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Washington D. C., USA: IEEE Press, 2012: 289-297.

[3] XI X, KEOGH E, SHELTON C, et al. Fast time series classification using numerosity reduction[C]//Proceedings of the 23rd International Conference on Machine Learning. New York, USA: ACM Press, 2006: 1033-1040.

[4] SWITONSKI A, JOSINSKI H, WOJCIECHOWSKI K, et al. Dynamic time warping in classification and selection of motion capture data[J]. Multidimensional Systems and Signal Processing, 2018(6): 1-32.

[5] LI S C X, MARLIN B M. Classification of sparse and irregularly sampled time series with mixtures of expected Gaussian kernels and random features[C]//Proceedings of the 31st Conference on Uncertainty in Artificial Intelligence. New York, USA: ACM Press, 2015: 484-493.

[6] TIRUNAGARI S, BULL S, POH N. Automatic classification of irregularly sampled time series with unequal lengths: a case study on estimated glomerular filtration rate[C]// Proceedings of the 26th International Workshop on Machine Learning for Signal Processing. Washington D. C., USA: IEEE Press, 2016: 1-6.

[7] SHEN Y, TINO P, TSANEVA-ATANASOVA K. Classification of sparsely and irregularly sampled time series: a learning in model space approach[C]// Proceedings of International Joint Conference on Neural Networks. Washington D. C., USA: IEEE Press, 2017: 3696-3703.

[8] FULCHER B D, JONES N S. Highly comparative feature-based time-series classification[J]. IEEE Transactions on Knowledge and Data Engineering, 2014, 26(12): 3026-3037.

[9] PAJARES R G, BENITEZ J M, PALMERO G S. Feature selection for time series forecasting: a case study[C]// Proceedings of the 8th International Conference on Hybrid Intelligent Systems. Washington D. C., USA: IEEE Press, 2008: 555-560.

[10] AHMED A A E, TRAORE I. Detecting computer intrusions using behavioral biometrics[C]//Proceedings of the 3rd Annual Conference on Privacy, Security and Trust. Washington D. C., USA: IEEE Press, 2005: 91-98.

[11] OUYANG Zhiyou, SUN Xiaokui. Human-machine behavior recognition for CAPTCHA based on gradient boosting model[J]. Netinfo Security, 2017(9): 143-146. (in Chinese)
欧阳志友，孙孝魁. 基于梯度提升模型的行为式验证码人机识别[J]. 信息网络安全，2017(9)：143-146.

[12] XIE Miao, LIU Linlan. Mouse trajectory recognition method based on naive Bayes[J]. Information & Communications, 2018(9): 30-32. (in Chinese)
谢苗，刘琳岚. 基于朴素贝叶斯的鼠标轨迹识别方法[J]. 信息通信，2018(9)：30-32.

[13] ZHANG Zhiteng, LIU Linlan. Mouse trajectory recognition method based on gradient boosted decision tree[J]. Information & Communications, 2018, 189(9): 22-24. (in Chinese)
张志腾，刘琳岚. 基于梯度提升决策树的鼠标轨迹识别方法与研究[J]. 信息通信，2018，189(9)：22-24.

[14] MENG Guangting, WANG Hong, LIU Haiyan. Mouse trajectory recognition method based on parallel voting decision tree and semi-supervised learning[J]. Journal of Chinese Computer Systems, 2018, 39(9): 2050-2055. (in Chinese)
孟广婷，王红，刘海燕. 融合并行投票决策树和半监督学习的鼠标轨迹识别方法[J]. 小型微型计算机系统，2018，39(9)：2050-2055.

[15] "China university computer contest-big data challenge"[EB/OL]. [2019-10-22]. http://bdc.saikr.com/bdc. (in Chinese)
"中国高校计算机大赛-大数据挑战赛"[EB/OL]. [2019-10-22]. http://bdc.saikr.com/bdc.

[16] BREIMAN L. Random forests[J]. Machine Learning, 2001, 45(1): 5-32.

[17] BLUM A, MITCHELL T. Combining labeled and unlabeled data with co-training[C]//Proceedings of the 11th Annual Conference on Computational Learning Theory. New York, USA: ACM Press, 1998: 92-100.

[18] PAGLIOSA L D C, MELLO R F D. Semi-supervised classification on positive and unlabeled problems using cross-recurrence quantification analysis[J]. Pattern Recognition, 2018, 80: 53-63.

[19] ZHANG Yan, WU Baoguo, LÜ Danju, et al. Active learning algorithm based on Tri-training[J]. Computer Engineering, 2014, 40(6): 215-218. (in Chinese)
张雁，吴保国，吕丹桔，等. 基于Tri-training的主动学习算法[J]. 计算机工程，2014，40(6)：215-218.

[20] GEURTS P, ERNST D, WEHENKEL L. Extremely randomized trees[J]. Machine Learning, 2006, 63(1): 3-42.

编辑 刘继娟