doi :10.3969/j.issn.1671-1122.2017.09.033

基于梯度提升模型的行为式验证码人机识别

欧阳志友 [1,2]，孙孝魁 [1,2]

（1. 南京邮电大学先进技术研究院，江苏南京 210023；2. 南京邮电大学自动化学院，江苏南京 210023）

摘　要：通过使用非正常手段模拟人类操作行为，绕过验证码系统，黑客工具就可以向系统后台发起批量请求，实现对系统的攻击，从而给系统的正常运行带来很大的风险，轻则影响系统运行，重则产生巨大的经济损失。而传统的验证码方法，在易用性和人机识别率方面都存在不足，行为式验证码应运而生。文章提出了一种基于行为式验证码的行为轨迹信息来构建特征工程，并运用梯度提升模型来进行人机行为识别的方法，在 10 万真实的行为轨迹样本上可以获得 90% 以上的识别准确率。

关键词：梯度提升；验证码；机器学习；人机识别

中图分类号：TP399　　**文献标识码：**A　　**文章编号：**1671-1122（2017）09-0143-04

中文引用格式：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.

英文引用格式：OUYANG Zhiyou, SUN Xiaokui. Human-machine Behavior Recognition for CAPTCHA Based on Gradient Boosting Model[J]. Netinfo Security, 2017(9):143-146.

Human-machine Behavior Recognition for CAPTCHA Based on Gradient Boosting Model

OUYANG Zhiyou[1,2], SUN Xiaokui[1,2]

(1. *Institute of Advanced Technology*, *Nanjing University of Posts and Telecommunications*, *Nanjing Jiangsu* 210023, *China*; 2. *School of Automation*, *Nanjing University of Posts and Telecommunications*, *Nanjing Jiangsu* 210023, *China*)

Abstract: By using the abnormal means to simulate human behavior operation and bypass the CAPTCHA system, hacking tools can then sent a large batch of requests to the background system to achieved the hacking goals, which may bring to big risk of delay response of system operation, or even produce huge economic losses. However, the traditional verification code method has shortcomings in both ease of use and man-machine recognition rate. In this paper, a new behavior trajectory of the CAPTCHA system based feature engineering, with utilizes the gradient boosting models, for human-machine behavior recognition is proposed. Performance in 100000 samples of real CAPTCHA system can obtain a more than 90% recognition accuracy.

Key words: gradient boosting; CAPTCHA; machine learning; human-machine recognition

收稿日期：2017-8-1

基金项目：国家自然科学基金重点项目 [61533010]；南京邮电大学实验室工作研究课题重点项目 [2014XSG03]

作者简介：欧阳志友（1982—），男，湖南，实验师，博士研究生，主要研究方向为机器学习、电力大数据分析；孙孝魁（1991—），男，河南，硕士，主要研究方向为机器学习和电力负荷预测与分析。

通信作者：欧阳志友 netivs@qq.com

0 引言

验证码是用来区分操作对象是人类还是机器的一种技术，它具有易操作、设计简单和传输数据小等特点，在拦截计算机自动化程序大批量的恶意行为方面，具有非常好的效果[1]。验证码的应用领域主要有阻止垃圾广告信息、保护网站注册和在线投票系统等[2]，在国内排名前 100 名的论坛中有超过 60% 的论坛在注册、登录或发帖部分采用验证码技术[3-4]，这个占比将会逐年增长。然而，在具体的应用中，不法分子可以运用模式识别等方法自动辨别验证码，攻克验证码这道关卡，实施自己的攻击行为。

常用的验证码主要有随机式和行为式验证码两种，其中行为式验证码又可以分为点触式和拖动式等[5]。随机式验证码主要是通过用户输入图片中的字母、数字、汉字等进行验证，是当前大多数网站采用的验证形式，其界面如图 1 所示。随机式验证码具有简单易操作、人机交互性较好等特点，但其安全系数较低，容易被破解。

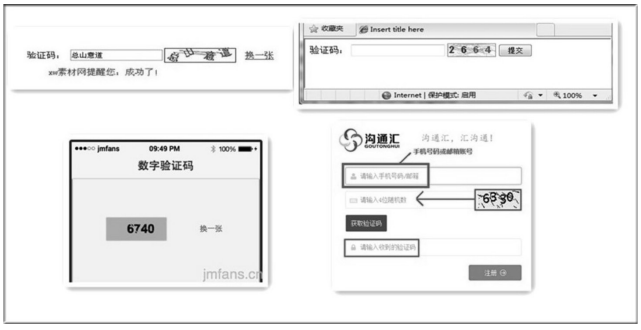

图 1 随机式验证码

点触式验证码是按要求点击其中一张或者多张图片，借用万物识别的难度阻挡机器，但是对图片、图库和技术的要求非常高。拖动式验证码是按照要求将备选碎片直线滑动到正确的位置，主要特点是操作简单，体验好，而且借助后台基于行为分析的机器学习模型，可以较好的防止破解，因此获得了更多的关注。对于随机式和点触式验证码来说，可以采集的人的行为信息不多，特征不够明显，其他的辅助功能也不太容易加入，只有在自身的验证码技术上取得突破；而拖动式验证码在输入验证码的过程中，需要鼠标拖动图标等一系列操作，存在大量的用户行为信息，因此可以使用机器学习方法来进行人机识别。拖动式验证码的界面如图 2 所示。

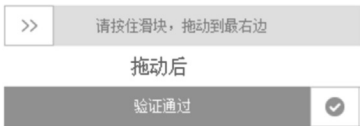

图 2 拖动式验证码

验证码的设计者与验证码的破解者的博弈一直都没有停止过，新的验证码的产生通常是旧的验证码被攻克或者旧的验证码在易用性或识别准确率方面存在不足，而新验证码的产生同时也促进了新的破解技术的诞生[6]。所以，我们除了要在验证码自身的技术上取得突破，还要在此基础上做一些辅助的工作。当验证码被攻克以后，要使验证码发挥作用，就必须在验证码输入完成以后，后台根据用户的一系列操作判断出该用户是人还是机器。当用户打开网页或者客户端之后，后台就开始采集用户的行为信息（鼠标轨迹），然后，用户进行业务操作（如登录、注册等），并且输入验证码完成之后，后台根据采集过来的行为信息利用算法判断该用户是人还是机器，如果判断是人且验证码输入正确，则允许访问；如果判断是机器，则直接拒绝。

在机器学习领域中，人机识别是一个典型的以人和机器为标签的二分类问题[7]。目前，主流的二分类算法有决策树（Decision Trees）[8]、随机森林（Random Forest）[9]、提升法 (Boosting)[10]、Logistic 回归算法[11] 和朴素贝叶斯法（Naive Bayesian）[12] 等。决策树算法适用于数据较少的训练集且具有很强的解释性，但是决策树算法的鲁棒性太差，训练数据微小的变化会导致决策树逻辑较大的变化。随机森林和提升法都属于集成算法，泛化能力强，对噪声数据不敏感，但是有时会出现过拟合的现象。Logistic 回归算法对 Y 属于哪一类的概率进行建模，可以筛选对建模影响最大的特征，但是对训练数据的规模要求比较大，且训练的时间比较长。朴素贝叶斯法是一种基于概率原则分类的学习算法，能处理好缺失数据和噪声数据，对大量数据和少量数据都适用，但是过分依赖条件独立性的假设，而这一假设往往是不成立的。

有鉴于此，本文通过构建一系列描述行为式验证码轨迹信息的特征，并利用梯度提升决策树算法对人机行为进行二分类识别，并获得了 90% 以上的准确率，可满足行为式验证码的使用需求。

1 行为式验证码人机识别方法

在行为式验证码中，展现给用户的图形或操作按钮只是起到初步的防护作用，其更主要的目的是要引导用户按照规范完成特定的行为，然后利用机器学习算法对行为轨迹进行建模，从而对行为进行综合判断，实现人机行为识别。用户完成行为式验证码的过程，其行为轨迹数据格式通常如表 1 所示。

表 1 用户行为轨迹数据格式

列名	格式	含义	示例
user_id	string	用户编号	1
action_time	int64	相对第一次操作的采样时间（毫秒数）	1734
action_x	int	采样时鼠标的 x 坐标	732
action_y	int	采样时鼠标的 y 坐标	623

从用户行为轨迹数据格式中可以看出，其采集的是从用户第一次操作行为式验证码开始的所有的鼠标位置，从而记录了用户在完成行为式验证码的验证过程中的鼠标轨迹。典型的机器行为的鼠标轨迹和人工行为的鼠标轨迹分别如图 3 a）和 3 b）所示，其中轨迹线之外独立的点为行为式验证码的目标点的坐标（这里的目标点的 y 坐标做了混淆，部分行为式验证码采集的数据中没有目标点坐标）。通常来说，行为式验证码对鼠标轨迹的采集时间要具有随机性，增加被破解的难度。

行为式验证码的核心问题是利用机器学习方法，基于用户鼠标轨迹数据来实现人机识别，这个识别过程又可以分为特征工程和机器学习模型两个主要的部分，其中特征工程完成从用户鼠标轨迹数据到机器学习模型输入数据之间的数据转换与处理工作，再通过机器学习模型完成人机识别。

1.1 特征工程

从行为式验证码的验证过程中采集的鼠标轨迹在时间、轨迹长度等方面存在很大的差异，无法直接用作机器学习算法的输入数据，因此需要先对这些数据进行处理，提取可以有效描述鼠标轨迹的特征信息作为模型的输入数据。为更好的描述鼠标轨迹行为，可以从时间、x 位移、y 位移、速度等多个方面提取特征，最终形成 60 多个特征，其中的主要特征如表 2 所示。

在具体提取特征前，需要先将每个用户的行为轨迹数据按用户编号和操作时间进行排序，然后计算两次相邻操作之间的坐标差、时间差和移动速度等。这些特征均可以通过 HIVE SQL 来实现，因此可以方便的在 Hadoop、阿里云数据平台（ODPS SQL）、腾讯 DIX 平台等大数据平台上进行实现，从而可以对大规模的行为式验证码进行人机识别。

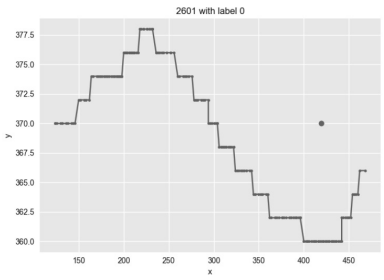

图 3 a）典型机器行为的鼠标轨迹

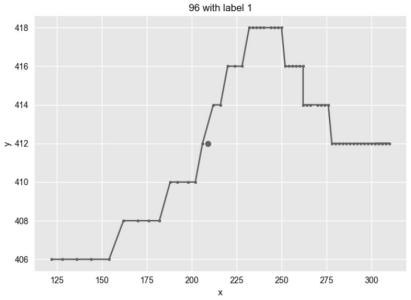

图 3b）典型人工行为的鼠标轨迹

表 2 行为轨迹特征

列名	格式	含义
max_action_x	int	鼠标轨迹的最大 x 坐标值
max_action_y	int	鼠标轨迹的最大 y 坐标值
max_x_offset	int	两次相邻采样的 x 最大间隔
max_y_offset	int	两次相邻采样的 y 最大间隔
x_back_cnt	int	相邻两次采样 x 位移为负数的次数
x_diff_entropy	double	相邻采样 x 位移的熵
x_diff_weighted_entropy	double	相邻采样 x 位移的加权熵
t_diff_entropy	double	采样时间间隔的熵
angel_entropy	double	相邻位移角度的熵
speed_entropy	double	移动速度的熵
action_cnt	int	采样次数
xy_duplicate_cnt	int	<x,y>坐标重复次数
total_t	int	完成全部操作的时间
diff_x_var	double	相邻采样间 x 位移的方差

1.2 梯度提升决策树模型

通过分析 2000 个标注好的行为式验证码的用户轨迹，可以发现有几种表现完全不一样的机器行为，一种比较好的方式是采用梯度提升决策树模型来进行人机识别分类预测。典型的梯度提升决策树模型的实现包括 GBDT（Gradient Boosting Decision Tree）、xgboost[13] 和 lightGBM[14] 等，这里选择用当前最流行的两种梯度提升决策树模型来具体实现：xgboost 和微软的 lightGBM，并使用了排序融合的集成学习方法，其人机识别模型的数据流程如图 4 所示。

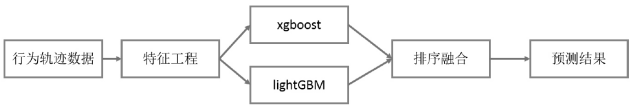

图 4 行为式验证码人机识别模型数据流程

考虑到机器行为具有多种完全不同的表现形式，并且不同的模型预测出来的概率（置信度）相差较大的情况，这里采用排序融合而非通常使用的基于预测概率进行融合的方式，即分别取 xgboost 模型预测为黑样本置信度最高的前 M 个（如前 19800 个）样本和 lightGBM 模型预测为黑样本置信度最高的 N 个（如前 16500 个）样本来进行去重合并，作为最终的预测结果。

2 性能分析

为检验基于梯度提升模型的行为式验证码人机识别的性能，使用了某互联网公司提供的来自真实拖拽式验证码的用户行为样本共计 10.2 万个，其中 400 个黑样本和 2600 个白样本作为二分类监督学习的训练样本。另外 10 万个为测试集，由该公司提供线上评测，对提交的预测结果给出得分。线上评测的评测函数为 $F=5PR\times100/(2P+3R)$，其中 P 为预测准确率，即预测为黑的数据（机器行为）中真正为黑的数量 / 判黑的数据总量（预测为机器行为的总数），R 为召回率，即预测黑的数据中真正为黑的数量 / 真实黑数据总量（真实的机器行为总数）。

在 python 语言中实现特征提取后，调用 xgboost 和 lightGBM 的 python 接口分别建模对 10 万个测试样本进行了预测，并对预测结果做了排序融合，其效果分别如表 3 所示。

表 3 梯度提升模型的预测效果

模型	提交数	得分	精度（%）	召回率（%）	说明
xgboost	15191	86.56	97.52	74.07	xgboost 预测为黑的所有样本
xgboost	20000	91.03	91.03	91.03	xgboost 预测概率前 2 万个
lightGBM	14472	85.5	98.56	71.32	lightGBM 预测为黑的所有样本
lightGBM	20000	86.98	86.98	86.98	lightGBM 预测概率前 2 万个
xgboost+lightGBM	15957	88	96.92	77.33	两个模型有一个预测为黑
xgboost+lightGBM	20000	89.34	89.34	89.34	两个模型预测概率均值前 2 万
xgboost+lightGBM	20005	91.04	91.03	91.06	xgboost 预测概率前 19800 和 lightGBM 预测概率前 16800

从表 3 中可以看出，基于梯度提升模型的预测结果精度和召回率均可以达到 91% 以上，可基本满足行为式验证码的实用要求。在模型融合方面，可以看出，使用概率均值的融合不如用排序融合的方式，但 lightGBM 得分偏低，导致模型融合的效果提升非常有限。

3 结束语

行为式验证码是近几年兴起的新型验证码，通过后台的机器学习模型来更好的避免验证过程被破解，并且可以很好的简化用户操作，提升验证的速度和效率，获得了广泛的应用。本文提出了一种基于梯度提升的行为式验证码的人机识别方法，并且给出了用户行为轨迹数据的特征提取方法和基于梯度提升决策树的机器学习模型，可以方便的移植到 spark、阿里巴巴的数加和腾讯的 DIX 等大数据平台上，从而实现大规模的行为式验证码人机识别。（责编 吴晶）

参考文献：

[1] 文晓阳，高能，夏鲁宁，等．高效的验证码识别技术与验证码分类思想 [J]. 计算机工程，2009, 35(8):186-188.

[2] 文伟平，郭荣华，孟正，等．信息安全风险评估关键技术研究与实现 [J]. 信息网络安全，2015（2）：7-14.

[3] 文晓阳，高能，荆继武．论坛验证码技术的安全性分析 [C]// 中国计算机学会计算机安全专业委员会，中国电子学会计算机工程与应用分会计算机安全保密学组．第 22 次全国计算机安全学术交流会，9 月 16-19,2007, 河北省张家界．北京：《信息网络安全》杂志社，2007:45-50.

[4] 张平，陈长松，胡红钢．基于分组密码的认证加密工作模式 [J]. 信息网络安全，2014（11）：8-17.

[5] HERNANDEZ C C J， BARRERO D F, RMORENO M D. Machine Learning and Empathy: the Civil Rights CAPTCHA[J]. Concurrency & Computation Practice & Experience, 2016, 28(4):1310-1323.

[6] KHEMCHANDANI R, SHARMA S. Robust Parametric Twin Support Vector Machine and Its Application in Human Activity Recognition[C] // Department of Computer Science and Engineering. Indian Institute of Technology. Proceedings of International Conference on Computer Vision and Image Processing, February 26-28, 2016, Roorkee, Uttarakhand, India. Singapore: Springer, 2017: 193-203.

[7] 苏涛．基于梯度提升树的行为式验证码的人机识别的研究 [D]. 武汉：华中师范大学，2016.

[8]KWAK N J, SONG T S. Android-Based Human Action Recognition Alarm Service Using Action Recognition Parameter and Decision Tree[J]. International Journal of Security & Its Applications, 2013, 7(4):277-286.

[9] GAN L, CHEN F. Human Action Recognition Using APJ3D and Random Forests[J]. Journal of Software, 2013, 8(9):412-423.

[10] MAZAAR H, EMARY E, ONSI H. Ensemble Based-Feature Selection on Human Activity Recognition[C] // Cairo University. Proceedings of the 10th International Conference on Informatics and Systems, May 9-11, 2016, Cairo, Egypt. America: ACM, 2016: 81-87.

[11] FAROOQ F, TANDON S, PARASHAR P, et al. Vectorized code implementation of Logistic Regression and Artificial Neural Networks to recognize handwritten digit[C] //Delhi Technological University. Power Electronics. Intelligent Control and Energy Systems (ICPEICES), July 04-06, 2016, Delhi, India. New York: IEEE, 2016: 1-5.

[12] FLACH P A, LACHICHE N. Naive Bayesian Classification of Structured Data[J]. Machine Learning, 2004, 57(3):233-269.