计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.190900107

一种新的设备指纹特征选择及模型构建方法

王 萌 丁志军
嵌入式系统与服务计算教育部重点实验室(同济大学) 上海 201804
上海市电子交易与信息服务协同创新中心(同济大学) 上海 201804
(1830799@tongji.edu.cn)

摘 要 近年来,随着移动互联网的快速发展,越来越多的业务从浏览器端转移到了移动端。但是,寄生在移动互联网上的黑色产业链也达到了泛滥的地步。设备指纹技术应运而生,即利用设备的特征属性为每个设备生成独一无二的标识。其间涌现了很多利用机器学习方法进行设备唯一性认证的策略,其中大部分方法注重于模型的建立,很少对特征选择部分展开深入研究,而特征选择直接关系到最终模型的性能。针对该问题,文中提出了一种新的设备指纹特征选择及模型构建方法(Feature Selection Based on Discrimination and Stability and Weight-based Similarity Calculation,FSDS-WSC),即根据不同设备的特征区分度和相同设备的特征稳定性选出最具价值的一些特征,并将这些特征的重要程度作为特征权重应用到模型建立的后续过程中。在真实场景中的6424台Android设备上,将FSDS-WSC与当今主流的其他特征选择方法进行了对比实验。结果表明,FSDS-WSC相比其他方法有了较大改进,设备唯一性认证的准确率达到了99.53%,证实了FSDS-WSC的优越性。

关键词:设备指纹;特征选择;相似度;权重;区分度;稳定性
中图法分类号 TP3-05

New Device Fingerprint Feature Selection and Model Construction Method

WANG Meng and DING Zhi-jun
Key Laboratory of Embedded System and Service Computing of Ministry of Education (Tongji University),Shanghai 201804,China
Shanghai Electronic Transactions and Information Service Collaborative Innovation Center (Tongji University),Shanghai 201804,China

Abstract In recent years,with the rapid development of mobile Internet,more and more businesses have moved from the browser to the mobile. But the black industry chain that is parasitic on the mobile Internet has reached the point of flooding. To solve this problem,the device fingerprint,that is,the use of the device's characteristic attributes to generate a unique identifier for each device came into being. Many algorithms based on machine learning methods for device uniqueness authentication have emerged,most of which focus on the establishment of models. Few of them have in-depth research on feature selection. However,feature selection is directly related to the performance of the final model. Aiming at this problem,this paper proposes a new device fingerprint feature selection and model construction method (FSDS-WSC),which is based on the feature discrimination of different devices and the feature stability of the same device to select some of the most valuable features. The importance of the selected features' weights is applied to the later model establishment. The FSFS-WSC is compared with other mainstream feature selection methods on 6424 Android devices in the real sence. The results show that FSFS-WSC has a great improvement compared with other methods,and the accuracy of device uniqueness authentication reaches 99.53%,which shows the superiority of FSFS-WSC.

Keywords Device fingerprint,Feature selection,Similarity,Weight,Discrimination,Stability

1 引言

随着计算机技术的飞速发展和移动互联网的兴起,尤其是互联网金融行业的迅猛崛起,寄生在移动互联网上的黑色产业链也达到了泛滥的地步。很多黑色产业通过虚假交易、伪造其他人身份、金融信用欺诈等方式骗取大量钱财。设备指纹技术应运而生,即利用设备的特征属性(包括硬件信息、软件信息、用户行为习惯)来为每个设备生成独一无二的标识符[1]。利用标识符,可以对设备进行有效的唯一性认证。

近些年来,移动端设备指纹主要经历了2.5代发展[2]。第一代设备指纹技术主要基于显性标识符即设备ID进行设备的跟踪。第二代设备指纹技术主要是最早由Eckersley[3]

到稿日期:2019-09-16 返修日期:2020-03-28
基金项目:国家自然科学基金面上项目(61672381);中央高校基本科研业务费专项资金重点项目(22120180508)
This work was supported by the National Natural Science Foundation of China (61672381) and Fundamental Research Funds for the Central Universities (22120180508).
通信作者:丁志军(zhujun_ding@outlook.com)

提出的针对浏览器的应用程序。但是随着移动设备的快速发展，在很多场景下，APP 应用程序已经逐渐取代了浏览器[4-5]，因此基于浏览器的设备指纹技术在当今并不适用。第 2.5 代设备指纹主要利用加速度计芯片中的细微差异导致的不同设备在同样静止状态下产生不同的间隙值，并足以用其区分不同设备。利用加速度传感器的研究可参考文献[6-9]。2015 年，Hupperich[10]等利用浏览器、系统、硬件等层面的设备信息，研究了一种近邻匹配方法来识别已知和未知的设备数据。其基本思想是：对采集到的设备信息基准库中的每一条设备信息的汉明距离进行计算。将最小的距离记为 d，若 d 小于给定阈值 T，则将该设备识别为距离为 d 的基准设备，并将该设备信息作为该设备的最新基准；否则，将该设备识别为新设备。该方案的主要不足是，没有在特征选择部分展开深入研究，即没有对这些特征进行选择，并且在相似度计算中没有为每个特征赋予权重，认为每个特征的贡献是相同的。

特征选择(Feature Selection)作为一种常见的降维方法，是近年来的研究热点之一。它指根据某一策略从原有特征空间选取部分特征，使得在模型中利用该特征子集有着较好的性能[11]。过滤法(Filter Method)是特征选择的主要方法之一，指不依赖于特定的学习算法进行特征选择。一种著名的过滤式特征选择方法是 Relief[12]，该方法设计了一个“相关统计量”来度量特征的重要性。其变体 Relief-F[13]，能够处理多分类问题。后来出现了结合统计学和信息论等多门学科的思想，并根据数据集的内在特性来评价每个特征的预测能力。常见的方法有皮尔森相关系数、信息增益等。这两种方法都通过评估特征与类别的关联程度考虑了特征的区分性。但是对于不同的非线性函数关系，这种方法给出的结果存在不一致的情况(由于是确定的函数关系，所以理论上结果应该都为 1)，即不具备普适性。这些特征选择方法的另一个缺点是仅在欧氏空间通过求方差考虑特征的稳定性，因此只适用于连续变量，对于大部分变量是类别变量的设备指纹类问题并不适用。通过对以上内容的研究和分析，本文主要贡献如下：

(1)利用 2011 年 David[14]提出的最大信息系数来衡量不同设备特征的区分度，并利用相同类别特征变化的频率来衡量设备特征的稳定性，通过综合设备的区分度以及稳定性提出一种基于区分度与稳定性的特征选择方案。

(2)根据特征选择的结果，为每个特征赋予一个特征权重。对上述近邻匹配方案中的相似度计算(汉明距离计算)部分进行改进。

(3)在真实场景中的 6 424 台 Android 设备上将 FSDS-WSC 与其他特征选择方法进行了对比实验验证。

本文第 2 节给出基于区分度与稳定性的特征选择方案以及改进的相似度计算方案；第 3 节描述使用的数据集以及模型在数据集上的实验效果；最后总结全文。

2 FSDS-WSC 模型

FSDS-WSC 综合考虑设备特征字段的区分度与稳定性，选出其中评分较高的特征，并将评分看作该特征的重要程度，以此作为后续设备唯一性认证中相似度计算部分的权重。FSDS-WSC 的整体框架如图 1 所示。

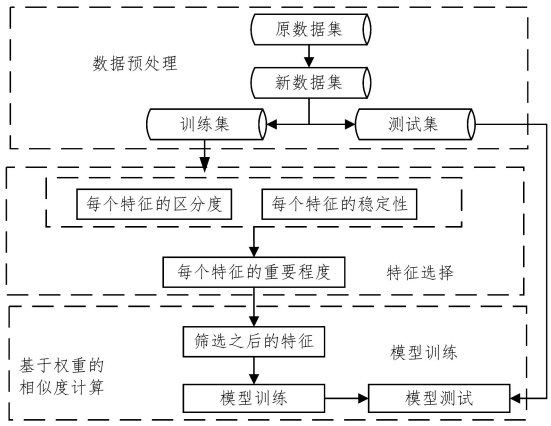

图 1 FSDS-WSC 整体框架

Fig. 1 FSDS-WSC overall framework

2.1 区分度

区分度指某特征与所属类别的关联程度。关联程度越高，则该特征的区分度越高。本文采用基于最大信息系数(Maximal Information Coefficient，MIC)的方法。MIC 能快速通过给不同类型的关联关系进行评估[15]，发现广泛范围的关系类型，成为了近几年应用金融领域风控中[16-17]较为先进的特征选择方法之一。MIC 可以检测各种类型的函数关系，适合在较大的数据集中去寻找两个变量之间的线性或复杂的非线性函数关系。其基本思想是：若两个随机变量存在某种线性或非线性关系，那么在这两个随机变量的散点图上可以按照某种方法画出一个网格，使得大多数点集中在网格的几个单元格内。该方法最大的优点是具有优秀的普适性(见表 1)[18]，即对任意的函数关系在无噪音影响时，其得分都为 1。其次，这种方法使得每个字段的区分度取值都位于[0，1]之间。在设备指纹场景下，每台设备可以产生很多条设备记录，并且伴随着时间的更新，会有新设备的接入，所以满足 MIC “大数据集”的特点。

表 1 不同函数关系下的多种测度

Table 1 Multiple measures under different functions

Function Type	Function	*MIC*	*Peterson*	*MI*
Linear	$y=x$	1.0	1.0	1.0
Quadratic	$y=4(x-1/2)^2$	1.0	−0.01	3.33
Sinusoidal	$y=\sin(7\pi x(1+x))$	1.0	−0.11	0.02

2.2 稳定性

由于用户行为习惯可能发生改变，即设备指纹具有易变性，因此除了需要考虑特征字段的区分度外，还需要考虑特征字段的稳定性。稳定性指相同类别中某一特征的变化情况。对于离散特征或类别特征，可以由该特征的变化频率来衡量稳定性；对于连续特征，可以利用方差来衡量稳定性。然后，通过特征的变化速率进行排序筛选。例如，在设备指纹场景中，某设备 A 共有 5 条设备记录，其中一个特征字段为“输入法”。为方便起见，不同的“输入法”采用不同的小写字母来表示。统计该 5 条记录中“输入法”随时间的变化情况，如表 2 所列。可以看出，第 4 条与第 5 条设备记录相较于前一条设备记录的输入法发生了变化。设备变化频率特征字段的变化频率越小，说明该特征的稳定性越高。

表 2　设备 A“输入法”的变化情况

Table 2　Changes of “input method” in device A

ID	InputMethod
1	a
2	a
3	a
4	b
5	c

在设备指纹数据集中，一台设备的不同设备记录中存在一些字段（例如剩余存储空间）的稳定性较差。即使某设备存在于基准库中，也可能由于这些字段的稳定性较差而使得该设备被判为新设备。因此，需要考虑每个字段的稳定性。

记某台设备 $S\in\Omega$ 的设备记录条数为 n，第 k 条记录与第 $k-1$ 条记录的某一个特征的变化情况记作 δ_k（$2\leqslant k\leqslant n$）。若这两条记录的对应特征相同，则 $\delta_k=0$；否则 $\delta_k=1$ 。统计该设备中每个特征变化的次数，记作 t_i（$1\leqslant i\leqslant total$），$total$ 为特征总数。t_i 及该设备的特征稳定性的计算式为：

$$t_i^S=\sum_{k=2}^{n}\delta_k^S \tag{1}$$

$$s_i^S=1-\frac{t_i^S}{n-1}(1\leqslant i\leqslant total) \tag{2}$$

在全部设备中，每个字段的平均稳定性 ms_i 的计算公式如下：

$$ms_i=\frac{1}{j}\sum_{q=1}^{j}s_i^*\ (*\in\Omega,j=|\Omega|) \tag{3}$$

2.3　特征的重要程度

基于特征的区分度与稳定性，每个特征的重要程度 $feature_w$ 按照下式进行计算：

$$feature_w=weight*m_w+(1-weight)*ms_w \tag{4}$$

其中，$weight$ 是权衡区分度与稳定性的参数，在一般情况下认为区分度的权重大于稳定性的权重；m_w 是第 w 个特征的最大信息系数；ms_w 是第 w 个特征的平均稳定性。

基于区分度与稳定性的特征选择方法如算法 1 所示。其中，$weight$ 为衡量区分度与稳定性的参数，若 $weight$ 大于 0.5 则认为区分度更重要，反之认为稳定性更重要。

算法 1　基于区分度与稳定性的特征选择算法

输入：训练集中的全部设备 $D=\{(a_1,b_1),\cdots,(a_n,b_n)\}$，其中 a_i 为设备信息，$b_i\in B$ 为设备类别，b_i^{num} 为该设备的设备记录条数，n 为记录总数（即训练数据有 n 条），j 为设备台数（即有 j 类）；total 为设备字段个数；Num 为最终选择的特征个数；Wgt 为权重

输出：最终选择的特征（i=1,2,3,…,Num）

1. for k ←1 to total do：
2. 计算特征字段 k 的最大信息系数 m_k；
3. 计算设备记录条数大于 1 的设备中每个特征字段的改变次数：$t_k^{b_i}=\sum_{s=2}^{b_i^{num}}\delta_s^{b_i}$；
4. 计算设备 b_i 中第 k 个字段的稳定性：$s_k^{b_i}=\frac{1-t_k^{b_i}}{b_k^{num}-1}$；
5. 计算字段 k 的平均稳定性：$ms_k=\frac{1}{j}\sum_{i=1}^{j}s_k^{b_i}$；
6. 计算每个特征的重要程度：$feature_w=Wgt*m_k+(1-Wgt)*ms_k$；
7. end for
8. 根据重要程度从大到小选出 Num 个特征 c_i（i=1,2,3,…,Num）。

2.4　基于特征权重的相似度计算

为方便给出具体的带有特征权重的相似度计算公式，首先定义距离向量和特征权重向量的概念。

定义 1（距离向量）　距离向量是一个 n 维向量 $\vec{\boldsymbol{A}}=(a_1,a_2,\cdots,a_n)$，其中 n 为设备的特征数。若两条设备记录的第 i（$1\leqslant i\leqslant n$）个特征相同，则 a_i 为 1，否则 a_i 为 0。

定义 2（特征权重向量）　特征权重向量是一个 n 维向量 $\vec{\boldsymbol{W}}=(w_1,w_2,\cdots,w_n)$，其中 n 为设备的特征数，w_i 为第 i 个特征的特征权重。

记某两台设备的距离向量为 $\vec{\boldsymbol{A}}$，特征权重向量为 $\vec{\boldsymbol{W}}$，则这两台设备的相似度计算公式为：

$$Similarity=\frac{\vec{\boldsymbol{A}}\cdot\vec{\boldsymbol{W}}}{n} \tag{5}$$

其中，· 为向量的内积运算。

3　实验设置与结果

3.1　实验数据集

本实验中使用的数据集是真实场景下的含有 6 424 台 Android 设备的 46 680 条数据记录。未进行特征选择时的初始特征数为 31 个。这些字段的数据类型包括整型、浮点型、布尔型、枚举型和字符串型等。每个设备字段的名称、数据类型如表 3 所列。

表 3　每个设备字段的名称、数据类型

Table 3　Name and data type of each device field

ID	Feature	Type
1	IMEI	String
2	MAC	String
3	brand	String
4	carrier	Enumeration
5	cpuCount	Integer
6	cpuHardware	String
7	breakflag	Boolean
8	cpuSerial	String
9	cpuSpeed	String
10	IMSI	String
11	meid	String
12	model	String
13	resolution	String
14	sim	String
15	wfmac	String
16	cpuABI	String
17	leftDisk	Float
18	leftMemory	Float
19	totalDisk	Float
20	totalMemory	Float
21	totalStorage	Float
22	country	Enumeration
23	defaultBrowser	String
24	inputMethod	String
25	inputMethodVersion	String
26	ip	String
27	language	Enumeration
28	networkType	Enumeration
29	phoneNum	String
30	IsSimulator	Boolean
31	timeZone	String

由表 3 可见，大部分设备的特征字段均为离散类型(String)，计算设备间的相似度比较耗时。于是，我们按照哈希散列[19]的方法将离散值数值化并划分为训练集和测试集。具体而言，将只有一条设备记录的设备放在测试集中，并随机选取一部分设备的全部设备记录作为新设备(相对于训练集中的设备而言)放在测试集中；其余设备根据登录时间，按照一定比例将每台设备的设备记录中登录时间早的放在训练集中，其余设备放在测试集中，使得最终训练集与测试集的比例为 6∶4。数据集划分结果如图 2 所示。

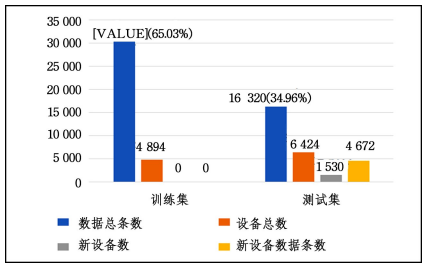

图 2 数据集划分结果

Fig. 2 Data set partitioning result

实验电脑的基本配置如下：处理器为 2.6 GHz Intel Core i7，内存 16 GB，操作系统为 Window10 64 位，编程环境为 python3.7。

3.2 实验指标与结果

在实验中，为了对比不同特征选择方法对性能的影响，采用控制变量的手段，即选择出来的特征个数与实验数据集等的设置均相同，只有特征选择方法不同，将皮尔森相关系数、信息增益两种特征选择方法作为对比。采用准确率(Accuracy)、错误接受率(False Accept Rate，FAR)、错误拒绝率(False Reject Rate，FRR)作为实验指标。混淆矩阵如表 4 所列。

表 4 混淆矩阵

Table 4 Confuse matrix

	预测为新设备	预测为旧设备
实际为新设备	TP	FN
实际为旧设备	FP	TN

(1)准确率：被正确分类的设备数占全部设备数的比例。准确率的计算公式如下：

$$Accuracy=\frac{TP+TN}{TP+TN+FP+FN} \tag{6}$$

(2)错误接受率：被误判为已有设备的样本数占所有新设备加上本身为旧设备但被判为其他旧设备的比例。错误接受率的计算公式如下：

$$FAR=\frac{FP}{TN+FP} \tag{7}$$

(3)错误拒绝率：被误判为新设备的已有设备的样本数占全部已有设备的比例。错误拒绝率的计算公式如下：

$$FRR=\frac{FN}{TP+FN} \tag{8}$$

按照算法 1，不同特征的最大信息系数以及稳定性如表 5 所列。

表 5 不同特征的最大信息系数以及稳定性

Table 5 *MIC* and *stability* of different features

ID	Feature	*MIC*	*Stability*
1	sim	1.0	0.987
2	imsi	1.0	0.987
3	imei	1.0	0.992
4	resolution	0.997	1.000
5	model	0.994	1.000
6	totalStorge	0.993	0.996
7	totalDisk	0.993	0.996
8	totalMemory	0.991	0.979
9	brand	0.989	1.000
10	mac	0.988	0.981
11	inputMethod	0.961	0.998
12	inputMethodVersion	0.946	0.967
13	leftDisk	0.944	0.01
14	cpuHardware	0.937	0.951
15	cpuSpeed	0.905	0.842
17	meid	0.801	0.996
18	phoneNum	0.634	0.993
19	carrier	0.605	0.992
20	wfmac	0.564	0.724
21	leftMemory	0.543	0.003
22	cpuSerial	0.475	1.000
23	cpuABI	0.475	1.000
24	ip	0.426	0.516
25	cpuCount	0.388	1.000
26	networkType	0.261	0.780
27	country	0.151	0.996
28	timeZone	0.0	1.000
29	IsSimulator	0.0	1.000
30	language	0.0	1.000
31	breakflag	0.0	1.000

若采用只考虑区分度的特征选择时，需要选择的特征个数可以根据实际场景从高到低选取若干个。为不失一般性，我们首先将 31 个不同特征的 *MIC* 从大到小进行排列(见图 3)，并计算前 30 个特征最大信息系数的差分(difference)。第 i 个特征的差分 k_i 的计算公式如下：

$$k_i=mic_{i+1}-mic_i(1\leqslant i\leqslant 30) \tag{9}$$

其中，mic_{i+1} 代表第 $i+1$ 个特征的 *MIC*，mic_i 代表第 i 个特征的 *MIC*。差分绝对值的物理意义指不同 *MIC* 下降的速率。

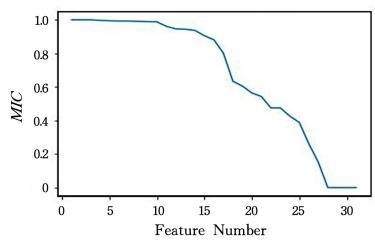

图 3 不同特征的最大信息系数

Fig. 3 MIC of different features

经计算得到第 17 个特征的差分绝对值最大，因此选择前 17 个特征进行实验。其他特征方法也选择 17 个特征。

在本文 FSDS-WSC 方法中，根据经验，我们得出区分度的优先级大于稳定性的优先级。因此，*weight* 选择 0.8。计算得到每个特征的重要程度如表 6 所列。

表 6　每个特征的重要程度

Table 6　Importance of each feature

ID	Feature	*Weight*
1	imei	0.998
2	resolution	0.998
3	imsi	0.997
4	Sim	0.997
5	model	0.995
6	totalDisk	0.994
7	totalStorage	0.994
8	brand	0.991
9	totalMemory	0.989
10	mac	0.987
11	inputMethod	0.968
12	inputMethodVersion	0.950
13	cpuHardware	0.940
14	defaultBroswer	0.903
15	cpuSpeed	0.892
16	meid	0.840
17	leftDisk	0.757

在基于皮尔森相关系数[20]的特征选择方法中，得到了排名靠前的几个特征以及对应的皮尔森相关系数，如图 4 所示。

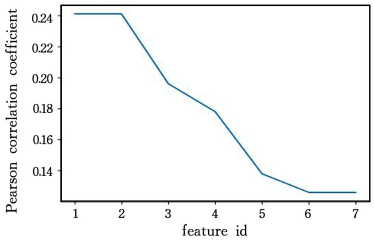

图 4　7 个皮尔森相关系数最大的特征

Fig. 4　Seven features with the maximum Pearson correlation coefficients

图 4 中，编号 1—7 分别为 leftMemory，totalMemory，wfmac，phoneNum，model，cpuABI 和 cpuSerial。而最相关的显性标识符如 imsi 和 imei 没有被检测出来。因此，皮尔森相关系数方法不适用于对被哈希数值化处理后的特征进行特征选择。

在基于信息增益[21]的特征选择方法中，排名前 17 位的特征归一化之后的信息增益如表 7 所列。

表 7　不同特征的信息增益

Table 7　Information gain of different features

ID	Feature	*IG*
1	imei	0.995
2	Sim	0.989
3	mac	0.989
4	imsi	0.989
5	model	0.976
6	totalStorage	0.965
7	totalDisk	0.965
8	totalMemory	0.960
9	inputMethodVersion	0.852
10	cpuSpeed	0.822
11	Ip	0.756
12	inputMethod	0.717
13	Leftdisk	0.693
14	leftMemory	0.690
15	Brand	0.689
16	Wfmac	0.657
17	resolution	0.612

每种方法的准确率、错误接受率和错误拒绝率如表 8 所列。

表 8　不同方法的性能指标

Table 8　Performance indicators of different methodsIDFeature

（单位：%）

ID	Feature Selection	Weight	*ACC*	*FAR*	*FRR*
1	无	无	97.62	10.28	1.45
2	MIC	无	99.29	0.72	0.71
3	IG	无	99.11	**0.60**	0.92
4	MIC	有	99.48	0.78	0.49
5	FSDS-WSC	有	**99.53**	0.78	**0.43**

由表 8 可知，在其他实验条件完全相同的情况下，相比实验 1，实验 2 和实验 3 通过不同的特征选择方法筛选出了最具有价值的特征，进而提高了模型性能，说明了特征选择在设备指纹模型构建方面的重要性。

另外，在特征选择方法完全相同（MIC）的情况下，相比实验 2，实验 4 在相似度计算中通过引入特征评分作为每个特征的权重，实验结果优于不加权重的方法。这也进一步说明了每个特征实际上对模型的性能贡献是不同的。

在其他实验条件完全相同的条件下，相比实验 4，实验 5 在特征选择中引入了稳定性的概念，实验结果优于只有区分度的结果，说明了特征的稳定性指标在特征选择中的重要性。

综上所述，本文提出的 FSDS-WSC 方法优于其他特征选择方法。

结束语　本文在原有的基于相似度计算的设备指纹模型上，提出了一种基于区分度与稳定性的特征选择方案和一种相似度计算改进方法。首先利用最大信息系数与设备字段更改频率从区分度与稳定性方面计算特征的重要程度，从而进行特征选择；之后将特征的重要程度作为特征权重引入到相似度计算（汉明距离计算）中。在包含 6424 台 Android 设备的 46680 条数据记录的真实数据集上的实验表明，本文方法在准确率上有显著提高，由此验证了其有效性。但是也应该注意到，FSDS-WSC 方法在阈值设置与特征权重设置方面是静态的，需要通过定期训练来更新阈值与特征权重，因此会花费较长的时间。如何在新数据到来时动态调节阈值以及特征权重，减少不必要的训练时间，提高模型的性能，是下一步需要研究的问题。

参 考 文 献

[1] BUJLOW T，CARELA-ESPAÑOL V，SOLEÉ-PARETA J，et al. Web tracking：Mechanisms，implications，and defenses[J]. Proc. of the IEEE，2017，105(8)：1476-1510.

[2] LIU J W，HUO Y M，WAN Y L. Review of equipment fingerprint research [C]// Proceedings of the 33rd National Computer Security Academic Exchange. 2018.

[3] ECKERSLEY P. How Unique is Your Web Browser? [C]// Privacy Enhancing Technologies，10th International Symposium (PETS 2010). Berlin：DBLP，2010.

[4] Mobile apps overtake PC Web usage in U. S. [EB/OL]. https://www. mendeley. com/catalogue/mobile-apps-overtake-pc-web-usage/.

[5] Percentage of all global web pages served to mobile phones from 2009 to 2018[EB/OL]. https://www. statista. com/statistics/

241462/global-mobile-phone-website-traffic-share/.

[6] BOJINOV H, MICHALEVSKY Y, NAKIBLY G, et al. Mobile Device Identification via Sensor Fingerprinting[J]. arXiv: 1408.1416.

[7] DEY S, ROY N, XU W, et al. AccelPrint: Imperfections of Accelerometers Make Smartphones Trackable[C]// Network and Distributed System Security Symposium. 2014.

[8] BALDINI G, AMERINI I, GENTILE C. Microphone identification using convolutional neural networks[J]. IEEE Sensors Lett., 2019, 3(7): 6001504.

[9] BALDINI G, STERI G. A survey of techniques for the identification of mobile phones using the physical fingerprints of the built-in compo nents[J]. IEEE Commun. Surveys Tuts., 2017, 19(3): 1761-1789.

[10] HUPPERICH T, MAIORCA D, MARC K, et al. On the Robustness of Mobile Device Fingerprinting: Can Mobile Users Escape Modern Web-Tracking Mechanisms? [C]// the 31st Annual Computer Security Applications Conference. ACM, 2015.

[11] CAI J, LUO J, WANG S, et al. Feature selection in machine learning: a new perspective[J]. Neurocomputing, 2018, 300: 70-79.

[12] KIRA K, RENDELL L A. The Feature Selection Problem: Traditional Methods and a New Algorithm[C]// Tenth National Conference on Artificial Intelligence. AAAI Press, 1992.

[13] KONONENKO I. Estimating attributes: analysis and extension of relief[C]// European Conference on Machine Learning. Berlin: Springer, 1994.

[14] RESHEF D N, RESHEF Y A, FINUCANE H K, et al. Detecting Novel Associations in Large Data Sets[J]. Science, 2011, 334 (6062): 1518-1524.

[15] WEN T, DONG D, CHEN Q, et al. Maximal Information Coefficient-Based Two-Stage Feature Selection Method for Railway Condition Monitoring[J]. IEEE Transactions on Intelligent Transportation Systems, 2019, 20(7): 2681-2690.

[16] PRAMESTI H, TALOMPO H R A. Determination of Credit Decision Attributes Using Maximal Information Coefficient [C]// International Conference on Information Technology Systems and Innovation (ICITSI). 2018.

[17] PONTIVEROS B B F, NORVILL R, STATE R. Monitoring the transaction selection policy of Bitcoin mining pools[C]// NOMS 2018 - 2018 IEEE/IFIP Network Operations and Management Symposium. IEEE, 2018.

[18] TAN Y Q, ZHANG X, LI Z, et al. Construction of Information Push Model Based on Maximum Mutual Information Coefficient [J]. Journal of Jilin University (Engineering Edition), 2018, 48(2): 558-563.

[19] Cryptographic hash funcation[EB/OL]. https://en.wikipedia.org/wiki/cryptographic_hash_ funcation.

[20] Pearson correlation coefficient[EB/OL]. https://en.wikipedia.org/wiki/Pearson_correlation_coefficient.

[21] COVER T M, THOMAS J A. Elements of Information Theory [M]. Wiley, 1991.

WANG Meng, born in 1996, postgraduate, is a member of China Computer Federation. His main research interests include machine learning, feature engineering.

DING Zhi-Jun, born in 1974, Ph.D, professor, Ph.D supervisor, is a member of China Computer Federation. His main research interests include service computing, formal method and intelligent system.