计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.181102118

基于深度森林与 CWGAN-GP 的移动应用网络行为分类与评估

蒋鹏飞　魏松杰
南京理工大学计算机科学与工程学院　南京 210094
(117106021927@njust.edu.cn)

摘　要　针对目前移动应用数目庞大、功能复杂,并且其中混杂着各式各样的恶意应用等问题,面向 Android 平台分析了应用程序的网络行为,对不同类别的应用程序设计了合理的网络行为触发事件以模拟网络交互行为,提出了网络事件行为序列,并利用改进的深度森林模型对应用进行分类识别,最优分类准确率可达 99.03%,并且其具有高精确率、高召回率、高 F1-Score 和低训练时间的特点。此外,为了解决应用样本数量有限且数据获取时间开销大等难题,还提出了一种使用 CWGAN-GP 的数据增强方法。与原始生成对抗网络相比,该模型训练更加稳定,仅需一次训练即可生成指定类别的数据。实验结果表明,在加入生成数据共同训练深度森林模型后,其分类准确率提高了 9%左右。

关键词:网络行为;应用分类;深度森林;流量分类;生成对抗网络
中图法分类号　P309

Classification and Evaluation of Mobile Application Network Behavior Based on Deep Forest and CWGAN-GP

JIANG Peng-fei and WEI Song-jie
School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing 210094, China

Abstract　In view of the problems that the large number and complex functions of mobile applications, and mixed with a variety of malicious applications, this paper analyzed the network behavior of applications for Android platform, and designed reasonable network behavior trigger events for different types of applications to simulate network interaction behavior. Based on the network event behavior sequence, the improved deep forest model is used to classify and identify applications. The optimal classification accuracy can reach 99.03%, and it has high accuracy, high recall rate, high F1-Score and low training time. In addition, in order to solve the problems of limited number of application samples and high time cost of data acquisition, a data enhancement method using CWGAN-GP was proposed. Compared with the original generative adversarial network, the training of the model is more stable, and the data of specified categories can be generated by only one training. The experimental results show that the classification accuracy is improved by about 9% after joining the generated data to train the deep forest model together.

Keywords　Network behavior, Application classification, Deep forest, Traffic classification, Generative adversarial network

1　引言

随着移动网络技术的发展和智能终端设备的普及,移动应用程序的数量呈爆炸式增长,恶意应用数量也迅速上升,其引发的危害包括资费消耗、隐私窃取、恶意扣费、远程控制等,这些都给移动终端用户造成了严重的负面影响。因此,移动应用的分类识别和恶意检测已经成为了当前亟待解决的关键问题。

目前,应用检测技术主要是基于应用程序的静态代码特征和动态运行特征,特征数据建模分析方法主要使用机器学习方法。静态代码特征分类大多数是通过反编译获取应用的申请权限[1-2]、API 调用列表[3-4]等特征来进行分类,但该方法可以通过重打包、代码混淆、动态更新等方法躲避基于静态特征的分类方法;基于动态运行特征方法是将应用程序放在特定的环境里运行以获取应用的流量消耗、电量消耗、内存使用和 CPU 占用率等运行特征,以及更复杂的特征,如 API 调用序列[5]、网络行为[4]等,但动态特征获取时间长,应用检测时间较慢;主流的机器学习分类方法主要有贝叶斯方法[6]、支持向量机[7]和 Autoclass 算法[8]等。然而,上述方法大都忽略了特征数据间的时序关系,因而它们的分类准确率都不太理想。

移动终端设备接入互联网能够产生大量的网络流量行为,且能够通过提取应用程序的网络交互流量特征识别应用程序的行为规律和特性,进而识别并评估应用程序的功能异常性。基于此,本文面向 Android 平台,使用不同的触发事件序列组合运行应用程序,收集应用产生的网络流量行为,将触发事件和其产生的网络流量行为组成网络事件行为序列,实

到稿日期:2018-11-16　返修日期:2019-04-30　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项目(61472189);赛尔网络下一代互联网技术创新项目(NGII20160105)
This work was supported by the National Natural Science Foundation of China (61472189), CERNET Innovation Project (NGII20160105).
通信作者:魏松杰(swei@njust.edu.cn)

现基于应用程序网络行为的分类。该方法可以实现跨操作系统的应用分类检测，具有很强的兼容性。

深度森林[9]是一种基于决策树的集成学习方法，有很强的表征学习能力，由多粒度扫描和级联结构两部分组成，因此也被称为多粒度级联森林(Multi-Grained Cascade Forest，gcForest)。级联结构用于表征学习，多粒度扫描用于提取数据的时序特征或空间特征。由于应用程序产生的一系列网络行为之间是有关联且相互影响的，因此本文将采用能够处理时序特征的深度森林方法对数据进行规约和建模。

通常情况下，同一类的应用网络行为大都具有相似性[10]，例如以酷狗音乐为代表的音频类应用主要是获取音乐、下载歌词和图片等。应用程序的行为一致性也是本文进行应用程序分类的基础，如果一个应用程序分类错误，很可能是该应用程序设置了其典型功能外的其他功能，这时应将其标记为风险应用，需要对其可靠性和安全性进行进一步测试和验证。因此，本文提出的模型可进一步扩展为恶意应用检测模型。

虽然应用市场中的应用数量足够多，但为了保证应用程序的可靠性和安全性，需要选取部分评分较高且下载数量大的应用作为正常应用样本。显然，正常应用样本的数量是有限的，而且应用程序的运行速度较慢，随着运行次数的增多，时间与资源的开销也会大大提升，但数据量过少又会对模型训练产生不利的影响。针对上述分析，本文提出了一种带条件和梯度惩罚的生成对抗网络(Conditional Wasserstein Generative Adversarial Network-Gradient Penalty，CWGAN-GP)进行数据增强，使用 CWGAN-GP 生成更多的相似数据，增加数据的多样性，并通过实验验证了生成数据能够有效地提升分类模型的准确性与健壮性。

2 网络事件行为序列

本节主要说明了应用的网络事件行为的获取方法，设计合理的应用触发事件让应用产生网络流量行为，然后对获取的网络流量行为数据进行清洗与处理。

2.1 网络事件行为序列的获取

本文使用 Monkeyrunner 工具编程实现应用程序的触发事件，在触发事件运行后获取其网络行为的流量数据，每次安装程序与运行触发事件都是独立的。本文从 Android 官方市场收集了 8 类应用，包括新闻类、视频类、音频类、管家类、拍照类、壁纸类、地图类和工具类，从 0 到 7 分别用数字 0－7 表示这 8 类应用的标签，即 0 表示新闻类，7 表示工具类。

由于这 8 类应用的功能和界面布局差异较大，因此需要设计合理的触发事件序列来更合理有效地触发各类应用的网络行为。本文针对每类应用的特点设计了合理的触发事件序列，考虑了触发事件之间的触发顺序和依赖关系，从而尽可能真实地模拟应用程序在使用过程中产生的网络行为。例如，全部应用程序的第一个触发时间必须是启动，视频类应用程序经常需要联网使用，管家类应用程序具有更大的可能性去触发清理内存行为。

本文共设计了 15 种触发事件，包括启动、使用网络、获取位置等。构建的网络事件行为序列模型如下：

$$((e_1,(s_1,r_1,c_1)),(e_2,(s_2,r_2,c_2)),\cdots,(e_i,(s_i,r_i,c_i)))$$

其中，e_i 表示触发事件标号，(s_i,r_i,c_i)表示触发事件发生后的网络行为，s_i 表示发送字节数，r_i 表示接收字节数，c_i 表示网络连接数，$(e_i,(s_i,r_i,c_i))$表示一个网络事件行为。当触发事件 e_i 发生时，它就会产生相应的网络行为数据(s_i,r_i,c_i)，所以 e_i 和(s_i,r_i,c_i)之间存在因果关系。同时，网络事件行为之间也存在时序关系，当一个触发事件完成后，下一个触发事件紧随发生，所以前一个网络事件行为会影响后续事件产生的网络行为。因此，本文采用能够处理时序特征的深度森林进行分类。

2.2 网络事件行为序列的处理

为了获取更有效的网络流量数据，本文进行了两方面的网络流量分析与清洗。

2.2.1 过滤无效的连接与会话

由于很多网络数据流的总发送字节数与总接收字节数都为 0，它们不但没有研究价值，反而容易造成虚高的总连接或会话数，因此对这类网络流量进行了剔除。

2.2.2 过滤广告产生的网络流量

由于大多数应用程序内会嵌入第三方广告库或在界面植入广告连接，模拟的点击交互事件很可能会点击到这些宣传广告，因此需要对捕获的网络流量再进行广告的清洗工作。为了过滤掉广告产生的网络流量，本文收集了一些较流行的第三方广告供应商的域名，并持续对这些域名进行解析以获取其 IP 地址集合。一段时间后，不难发现这些第三方广告的 IP 地址虽然是动态变化的，但仍存在一个常用的动态 IP 地址集合。本文归纳总结了广告 IP 地址集合，统计了不同类别应用中带有广告流量的比例，发现不同功能类别的应用访问广告网站的概率是不同的。其中，壁纸类应用的广告流量甚至超过了 20%，具体如图 1 所示。由此，我们可以通过这个动态 IP 地址集合进行常见广告流量的过滤处理。

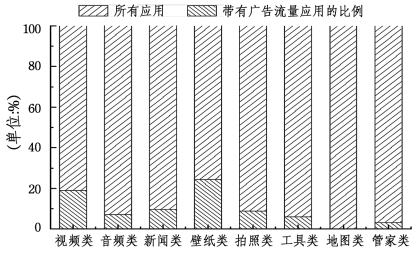

图 1 不同类别应用中带广告流量的应用比例

Fig. 1 Application ratio of advertising traffic in different categories of applications

本文为每类应用设计了 5 个触发事件，每个网络事件行为序列的长度为 20，但由于其发送字节数和接收字节数的数值较大，而且有很大的波动，会影响后面深度森林的训练，因此需要对其进行归一化处理。

$$s_i'=\lg(s_i),r_i'=\lg(r_i),c_i'=c_i \tag{1}$$

3 深度森林

3.1 深度森林简介

深度森林的级联森林部分和多粒度扫描部分是独立的，

可以在不使用多粒度扫描的前提下直接使用级联森林。

3.1.1 级联森林

深度森林的级联结构采用了类似深度神经网络的一层叠一层结构，即从前层输入数据，输出结果作为下层的输入。每一层都会生成长度为 C（类别数）的类别概率向量，如果一层有 N 个分类器，那么每个分类器生成的 C 个元素会拼接在一起，组成 $C\times N$ 个元素向量，再把源输入特征向量拼接上去，这样就组成了下一层的输入。

级联森林的每层由多个机器学习分类器组成，但需要注意的是，每层的分类器不能是同一种，比如一层全是由随机森林组成的，这是因为多样的结构对集成学习来说是非常重要的[11]。最后，每一个分类器都会输出一个类别概率向量，对这些类别概率向量求均值，值最大的一类就是最终预测结果。

级联森林在扩展一个新的层级后，整个级联的性能将在验证集上被估计，如果没有显著的性能增益，训练过程将终止。级联森林能够通过适当地终止训练来决定其模型的复杂度，这使得级联森林能够适用于不同规模的训练数据，而不局限于大规模训练数据。

3.1.2 多粒度扫描

受到循环神经网络和卷积神经网络的启发，gcForest 利用多粒度扫描方法增强级联森林，这使得 gcForest 在处理时序特征数据或空间特征数据时表现良好。

多粒度扫描使用滑动窗口扫描原始特征，并且可以通过使用多个尺寸的滑动窗口，最终的变换特征向量将包括更多的特征。多粒度扫描的过程如下：输入一个完整的 P 维样本，通过一个长度为 L 的采样窗口对其进行滑动采样，从而得到 $S=(P-K)/L+1$（L 表示滑动步长）个 K 维特征子样本向量；然后，每个子样本都用于完全随机森林和普通随机森林的训练中，并在每个森林中都获得一个长度为 C（类别数）的概率向量，这样每个森林会产生长度为 $S\times C$ 的表征向量；接着，把每层的 F 个森林的结果拼接在一起得到本层输出。假设有 n 个大小为 K 的滑动窗口，可以得到最终的输出维度为：

$$D=\sum_{i=1}^{n}\left[\left(\frac{P-K_i}{L}+1\right)\times C\times F\right] \tag{2}$$

最后，将多粒度扫描的输出用于级联森林的训练，以此来增强级联森林。

3.2 深度森林模型设计

针对网络行为序列的特点，设置了 3 个滑动窗口，滑动窗口大小分别为 8，12，20，滑动步长均为 1。另外，本实验特别增加了一种新的滑动方式，即让滑动采样的每 2 个元素之间间隔 I。例如，对于一个长度为 L 的序列样本，滑动窗口大小为 M（M 要能被 L 整除），这 2 个元素之间的间隔 $I=L/M$，这样就需要滑动 L/M 次，最后可以得到 L/M 个滑动切片，因此多粒度扫描的输出维度就为：

$$D'=D+L/M\times C\times F \tag{3}$$

本文实验中级联森林的每一层使用了 4 个分类器，包括 XGBoost[12] 和随机森林、完全随机森林和 Logistic 回归。gcForest 的总体结构参数如图 2 所示。

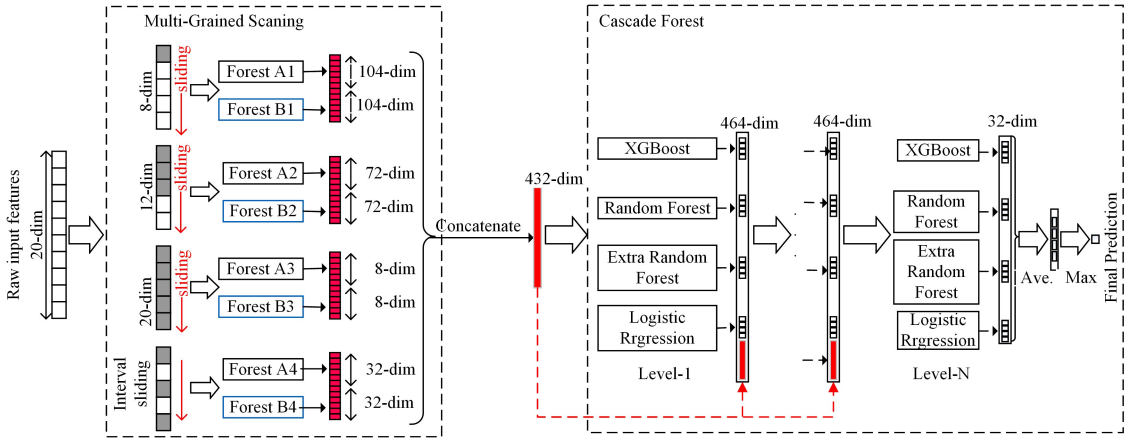

图 2 gcForest 结构

Fig. 2 Structure of gcForest

4 WGAN-GP

4.1 WGAN-GP 简介

相对于其他生成式模型，生成对抗网络[13]计算复杂度低，因此在图像方面具有非常好的应用效果。GAN 由生成模型（G）和判别模型（D）两部分组成，生成模型的输入是随机噪声 z，它将随机噪声 z 转化为符合真实数据分布的数据 $G(z)$。判别模型的输入是真实数据 x 和生成数据 $G(z)$，它用来判别输入数据是真实的还是生成的。

WGAN（Wasserstein GAN）[14] 从损失函数的角度对 GAN 易陷入局部最优解、训练不稳定等问题做了改进。WGAN 使用 Wassertein 距离去衡量生成数据分布和真实数据分布之间的距离。与 GAN 中用于衡量两个分布间距离的 JS 散度相比，Wassertein 距离的优越性在于，即便两个分布没有重叠，其仍然能够反映它们的远近。Wasserstein 距离定义如下：

$$W(p_r,p_g)=\inf_{\gamma\sim\prod(p_r,p_g)}\mathbb{E}_{(x,y)\sim\gamma}[\|x-y\|] \tag{4}$$

使用 Wassertein 距离需要满足很强的连续性条件——Lipschitz 连续性，WGAN 在整个样本空间上都施加了 Lipschitz 连续性，但 WGAN-GP[15] 指出这是没有必要的，只要重点抓住生成样本的集中区域、真实样本的集中区域以及夹在它们中间的区域即可。具体的做法是在判别器的损失函数上增加一个梯度惩罚项，先随机采样一对真假样本和一个服从均匀分布 U(0,1) 的随机数：$x_r\sim P_r$，$x_g\sim P_g$，$\varepsilon\sim Uniform(0,$

1)；然后在 x_r 和 x_g 的连线上随机插值采样：

$$\hat{x}=\varepsilon x_r(1-\varepsilon)x_g \tag{5}$$

把按照上述流程采样得到的 $\hat{x}$ 所满足的分布记为 $P_{\hat{x}}$，最终可得到判别器的损失函数为：

$$L(D)=-\mathbb{E}_{x\sim P_r}[D(x)]+\mathbb{E}_{x\sim P_g}[D(x)]+\lambda\mathbb{E}_{x\sim P_{\hat{x}}}[\|\nabla_x D(x)\|_p-1]^2 \tag{6}$$

WGAN-GP 提出了一种新的 Lipschitz 连续性限制手法——梯度惩罚，该限制手法解决了训练梯度消失和梯度爆炸的问题，比标准 WGAN 拥有更快的收敛速度，并能生成更高质量的样本，而且训练更加稳定，几乎不需要调参。

4.2 CWGAN-GP 模型设计

GAN 是一种无监督的学习方式，导致生成器比较自由，训练好的生成器不可控制，无法得知生成数据的类别。受到条件生成对抗网络的启发，本文给 WGAN-GP 的生成器和判别器均加入条件变量 y，提出了带条件的 CWGAN-GP，使用额外信息 y 对模型增加条件，这样就可以指导数据的生成过程。生成器和判别器的损失函数即变为：

$$L(G)=-\mathbb{E}_{x\sim P_{\hat{x}}}[D(x|y)] \tag{7}$$

$$L(D)=-\mathbb{E}_{x\sim P_r}[D(x|y)]+\mathbb{E}_{x\sim P_g}[D(x|y)]+\lambda\mathbb{E}_{x\sim P_{\hat{x}}}[\|\nabla_x D(x|y)\|_p-1]^2 \tag{8}$$

其中的条件变量 y 被设置为数据的标签。通过这种方法就可以同时训练多类数据，训练完成后，只需将噪声和指定类别的标签输入生成器就能得到相应类别的数据，这在很大程度上降低了训练时间开销。

本实验设计的 CWGAN-GP 网络结构如图 3 所示。生成器的输入由服从均匀分布的随机噪声 z 和数据标签组成，随机噪声的维度为 40。生成器包括 3 层全连接层，每层的节点数分别为 128，256，20，激活函数均使用 ReLu 函数。生成器生成数据之后，将生成数据再与标签组合输入到判别器中。判别器的输入由生成数据，真实数据和标签组成。判别器由 3 层全连接层组成，每层的节点数为 128，256，1，前两层使用 ReLu 激活函数，最后一层不使用激活函数。优化算法均使用 Adam 算法。

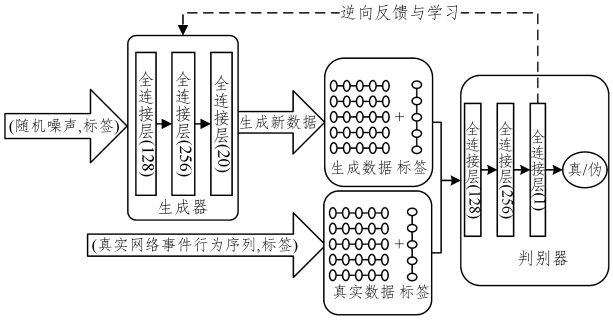

图 3 CWGAN-GP 模型结构

Fig. 3 Structure of CWGAN-GP

5 实验结果与分析

5.1 实验环境

本实验均在内存为 16 GB，处理器为 Intel(R)Core(TM) i7-7700 CPU 3.6 GHz，显卡为 Nvidia GeForce GTX 1070Ti 的台式机上操作。编程语言使用 Python2.7。gcForest 模型的实现基于机器学习框架 Scikit-learn(0.18.0)，CWGAN-GP 模型的搭建基于深度学习框架 Tensorflow(1.2.0，GPU 版本)。

5.2 实验环境

本实验从 Android 应用市场收集了共 559 个下载量普遍很高的 8 类应用程序，包括新闻类、管家类、工具类、视频类、壁纸类、拍照类、地图类、音频类，它们具有较高的实验价值与研究意义，具体数量如表 1 所列。

表 1 各类应用程序的分布

Table 1 Distribution of various applications

应用类别	个数	示例
新闻类	82	tengxun. news
管家类	44	anquan
工具类	69	huawei. moilenotes
视频类	91	youku. video
壁纸类	62	360. wallpaper
拍照类	78	pitu. boxxcam
地图类	58	com. tigerknows
音频类	75	com. kuwo. android
总数	559	

获取应用程序后，通过第 2 节介绍的网络流量特征的获取、清洗与提取方式得到各类应用程序的网络行为特征，然后将它们构建成网络事件行为序列。每个应用程序运行 10 次不同场景事件组合，一共重构了 5590 条网络事件行为序列。其中，4472 条数据作为 CWGAN-GP 生成模型的原始数据，用来生成新的相似数据；其余的 1118 条数据作为 gcForest 分类模型的测试数据。

5.3 CWGAN-GP 生成结果与分析

生成器和判别器的学习率均设置为 0.0008，λ 设置为 35，批次大小设置为 200。判别器和生成器的损失值变化情况如图 4、图 5 所示。

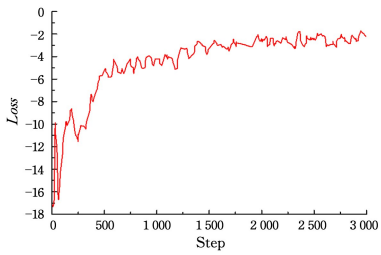

图 4 判别器的损失值

Fig. 4 Loss of discriminator

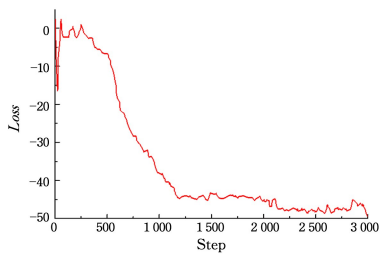

图 5 生成器的损失值

Fig. 5 Loss of generator

由图 4、图 5 可以看出，判别器的损失值在经过 250 轮训练后逐渐上升，最终越来越接近于 0，这表明判别器已经无法区分数据的来源。生成器的损失值不断下降，表明生成数据

与真实数据的相似度越来越高，导致判别器无法判断数据的来源。由于添加了条件限制和梯度惩罚，本文提出的 CWGAN-GP 在训练过程中没有出现像原始 GAN 那样训练不稳定的情况，并且其收敛速度也是非常快的。

本实验为每类应用的网络行为序列生成了 2000 条数据，共生成 16000 条新的网络事件行为数据，再加上原来的 4472 条真实数据，共 20472 条数据用于对 gcForest 分类模型进行训练，测试数据仍然用原来的 1118 条真实数据。

5.4 gcForest 分类结果与分析

为了验证本文所提出的网络事件行为序列数据生成模型的有效性，首先使用原始的 4472 条真实数据作为训练数据进行分类测试，然后使用原始的 4472 条真实数据加上逐步增加的生成数据进行训练。为每类应用的网络事件行为序列分别添加 0，300，600，900，1200，1500，1800，2000 条生成数据的平均测试分类准确率如表 2 所列。

表 2 逐步添加生成数据的平均测试分类准确率

Table 2 Incrementally adding average test classification accuracy of generated data

样本数量	测试准确率/%
4472	90.05
6872	94.19
9272	96.45
11672	98.06
14072	99.03
16472	97.52
18872	96.65
20472	96.24

从表 2 可以看出，当仅用原始的 4472 条真实数据时，测试集上的分类准确率为 90.05%。随着训练数据的不断增加，分类准确率也在不断上升，当每类应用添加 1200 条生成数据时，分类准确率最高可以达到 99.03%。此后再添加生成数据，分类准确率不升反降，这表明此时的生成数据已经过于泛化，训练数据的分布规律过于复杂，gcForest 模型的分类准确率已经不能再提高了。

考虑到级联森林中的随机森林分类器有一定的随机性，为了验证模型分类的稳定性，为每类应用的网络事件行为序列添加 1200 条生成数据，重新进行了 5 次实验，并且增加了精确率(Precision)、召回率(Recall)和 F1-Score(精确率和召回率的调和平均值)作为评价指标，实验结果如表 3 所列。

表 3 gcForest 模型的分类结果

Table 3 Classification results of gcForest model

(单位：%)

实验序号	精确率	召回率	F1-Score	准确率
1	99	99	99	98.71
2	98	98	98	98.39
3	97	97	97	97.45
4	99	99	99	99.03
5	98	98	98	98.39

由表 3 可知，使用真实数据与 CWGAN-GP 生成数据作为训练集的模型在测试集上的平均分类准确率最高可达 99.03%，平均分类准确率在 98.39%左右，并且其具有非常高的精确率、召回率和 F1-Score。可见，gcForest 的性能非常稳定，分类准确率亦较高。

此外，gcForest 的输出为预测到每一类的概率，分别表示一条网络事件行为序列属于各个功能类别的概率，因此可以通过统计各类应用的网络事件行为序列预测到 8 个功能类别的概率，从而更清晰、明确地观察到各类应用的网络行为与其他类的区别。为了进一步展示同一功能类目下的正常应用样本的高相似性和分类模型的高判别概率，本实验选取了测试结果中管家类的网络事件行为序列的 gcForest 的输出情况并对其进行分析，统计结果如图 6 所示。

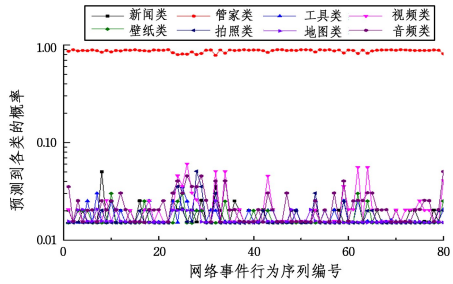

图 6 管家类应用 gcForest 模型的分类情况

Fig. 6 Classification of gcForest models in housekeeping application

从图 6 中可以发现，基本上每一条网络事件行为序列预测到管家类的概率都远高于其他类别的概率，并且判别概率都非常接近于 1。一方面，这说明了 gcForest 模型已经能够很好地学习与规约管家类应用的网络行为模式，对该类应用的判别概率非常高；另一方面，其也反映了管家类应用的网络行为与其他功能类别的行为有很大的区别，表明同类应用的网络行为具有很强的一致性。

根据应用的行为一致性理论，相同功能类别的不同应用程序之间，应用的行为模式具有相似性，产生的网络行为同样具有相似性。如果一个应用被错误分类或者其判别概率不高，很可能是因为该应用设置了其归属类别的典型功能之外的其他功能，此时它需要被标记为风险应用。本实验另外特别收集了 50 个疑似风险应用的地图类应用，gcForest 的分类结果输出如图 7 所示。

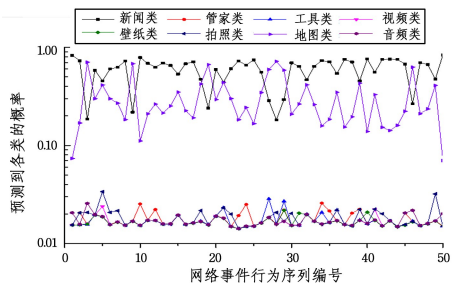

图 7 疑似风险应用的地图类应用的分类情况

Fig. 7 Classification of map applications for suspected risks applications

图 7 显示，疑似风险应用的分类情况非常杂乱，标记为地图类应用的行为模式和新闻类应用的行为模式非常相似，大多数地图类应用都被错误分类到了新闻类，即使被正确分到了声明的类别，分类器的判别概率也较低，与正常应用样本有很大的区别。这表明这些风险应用常常会推送很多消息，用户点击查看之后会产生类似新闻类的网络流量行为，这些疑似恶意应用的网络行为与正常应用样本有很大的差别，需要被标记为风险应用。因此，可将本文所提分类模型与方法进

一步扩展为恶意检测模型。

5.5 与其他方法的比较

为了验证本文所提方法的有效性，将其与其他方法进行了比较，包括决策树、随机森林、线性支持向量机(SVM)、BP 神经网络和长短时记忆网络(LSTM)。实验结果如表 4 所列。

表 4 各种方法的比较

Table 4 Comparison of methods

分类器	含生成数据		无生成数据	
	准确率/%	时间/min	准确率/%	时间/min
决策树	84.93	2	82.58	1
随机森林	86.12	2	82.58	1
线性 SVM	84.69	2	82.26	1
BP 网络	86.32	17	83.81	13
LSTM	92.77	28	85.72	21
深度森林	**99.03**	**10**	**90.05**	**7**

由表 4 可知，在添加生成的数据后，其他分类器的分类结果虽然有所改善，但它们最终的分类准确率仍然不高，准确率的提升也不明显，均低于本文的深度森林模型。由于 LSTM 也有选择记忆功能，因此其分类准确率也很高，可以达到 92.77%，但是 LSTM 是神经网络模型，训练起来非常耗时。在有 GPU 计算加速的情况下，LSTM 的训练时间仍然是仅靠 CPU 计算的 gcForest 的两倍。因此，无论从分类准确率还是时间性能上来看，本文提出的深度森林模型均优于常见的机器学习分类算法。

结束语 本文通过分析重构应用程序运行过程中基于不同触发事件组合产生的网络事件行为序列，进行应用程序的网络行为特征描述与规约，采用深度森林进行网络行为的训练与学习。所提模型对移动应用程序的分类准确率高达 99.03%，并可进一步扩展为恶意应用检测模型。鉴于真实数据样本数量不足的问题，本文又提出了一种高效的数据生成方法，即使用 CWGAN-GP 模型进行数据增强，根据现有的真实网络事件行为数据生成新的数据。实验表明，深度森林模型采用真实数据与 CWGAN-GP 模型的生成数据共同训练与学习，增加了数据的多样性，分类准确率提高了 9%左右。然而，虽然本文提出的 gcForest 模型可以扩展为恶意应用检测模型，但其没有实现在线应用检测功能，未来将在该方面进行研究，制定移动应用的判别概率阈值，实现基于本文提出的分类模型的恶意应用检测方法。

参 考 文 献

[1] GHORBANZADEH M, CHEN Y, MA Z M, et al. A neural network approach to category validation of Android applications [C]// 2013 International Conference on Computing, Networking and Communications (ICNC). San Diego: IEEE, 2013: 740-744.

[2] HAO H K, LI Z J, YU H B. An effective approach to measuring and assessing the risk of android application[C]// 2015 International Symposium on Theoretical Aspects of Software Engineering. Nanjing, China: IEEE, 2015: 31-38.

[3] WANG R, FENG D G, YANG Y, et al. Semantics-Based Malware Behavior Signature Extraction and Detection Method[J]. Journal of Software, 2012, 23(2): 378-393.

[4] WEI S J, YANG L. Android Malware Characterization Based on Static Analysis of Hierarchical API Usage [J]. Computer Science, 2015, 42(1): 155-158.

[5] CHUANG H Y, WANG S D. Machine learning based hybrid behavior models for Android malware analysis[C]// 2015 IEEE International Conference on Software Quality, Reliability and Security. Vancouver: IEEE, 2015: 201-206.

[6] MINN S, FU S, LV T. Algorithm for exact recovery of Bayesian network for classification[J]. Application Research of Computers, 2016, 33(5): 1327-1334.

[7] ESTE A, GRINGOLI F, SALGARELLI L. Support vector machines for TCP traffic classification[J]. Computer Networks, 2009, 53(14): 2476-2490.

[8] ZANDER S, NGUYEN T, ARMITAGE G. Automated traffic classification and application identification using machine learning[C]// The IEEE Conference on Local Computer Networks 30th Anniversary (LCN'05) l. Sydney: IEEE, 2005: 250-257.

[9] ZHOU Z H, FENG J. Deep forest: Towards an alternative to deep neural networks[J]. arXiv: 1702.08835, 2017.

[10] WANG J Y, XU M K, WANG H Y, et al. Automated Detection of the Inconsistence between App Behavior and Privacy Policy of Android Apps[J]. Journal of Frontiers of Computer Science and Technology, 2019, 13(1): 56-69.

[11] ZHOU Z H. Ensemble methods: foundations and algorithms [M]. CRC Press, 2012.

[12] CHEN T Q, GUESTRIN C. Xgboost: A scalable tree boosting system[C]// Proceedings of the 22nd Acmsigkdd International Conference on Knowledge Discovery and Data Mining. New York: ACM, 2016: 785-794.

[13] WANG K F, GOU C, DUAN Y J, et al. Generative Adversarial Networks: The State of the Art and Beyond [J]. Acta Automatica Sinica, 2017, 43(3): 321-332.

[14] ARJOVSKY M, CHINTALA S, BOTTOU L. Wasserstein gan [J]. arXiv: 1701.07875, 2017.

[15] GULRAJANI I, AHMED F, ARJOVSKY M, et al. Improved training of wasserstein gans[C]// Advances in Neural Information Processing Systems. Long Beach, USA: Neural Informaton Processing Sysmtes, 2017: 5767-5777.

JIANG Peng-fei, born in 1995, postgraduate, is not member of China Computer Federation (CCF). His main research interests include traffic analysis and deep learning.

WEI Song-jie, born in 1977, Ph.D, professor, is member of China Computer Federation (CCF). His main research interests include network security, network data analysis and monitoring, abnormal event detection and simulation.