基于多态特征的Android应用恶意行为检测系统

杨欢1） 张玉清1）2）  胡予濮1） 刘奇旭2）

1）（西安电子科技大学综合业务网理论及关键技术国家重点实验室 西安 710071）

2）（中国科学院大学国家计算机网络入侵防范中心 北京 100190）

**摘要** 目前针对未知的Android恶意应用可以采用数据挖掘算法进行检测，但使用单一数据挖掘算法无法充分发挥Android应用的多类行为特征在恶意代码检测上所起的不同作用，文中首次提出了一种综合考虑Android多类行为特征的三层混合系统算法THEA（Triple Hybrid Ensemble algorithm）用于检测Android未知恶意应用，首先，采用动静态结合的方法提取可以反映Android应用的恶意行为的组件、函数调用及系统调用类特征；然后，针对上述3类特征设计了三层混合系统算法THRA，该算法通过构建合适3类特征的最优分类器来综合评测Android应用的恶意行为；最后，基于THRA实现了Android应用恶意行为检测工具Androdect,并对现实中的1126个恶意应用和2000个非恶意应用程序进行检测，实验结果表明，Androdect能够利用Android应用的多类行为特征有效检测Android未知恶意应用，并且与其他相关工作对比，Acdrodect在检测准确率和执行效率上表现更优。

**关键词** 系统算法；Android应用；多态特征；恶意代码检测；行为分析；数据挖掘；智能手机；网路行为

**中图法分类号** TP393 **DOI号** 10.3724/SP.J.1016.2014.00015

**A Malware Behavior Detection System of Android Application Based on Multi-Class Features**

YANG Huan1) ZHANG Yu-Qing1)2) HU Yu-Pu1) LIU Qi-Xu2)

1) (State Key Laboratory Integrated Services Networks, Xidian University, Xi’ an 710071)

2)(National Computer Network Intrusion Protection Center, University of Chinese Academy of Sciences ,Beijing 100190)

**Abstract** At present, data mining algorithm is always used to detect unknown malicious applications of Android. While single data mining algorithm could not play the role of multi-class Android features in malware detection. For this problem, a Triple Hybrid Ensemble Algorithm (THRA) was first proposed ,which considered multi-class Android features. First, we combined the static and dynamic methods to extract three classes of Android features, which could reflect malicious behavior effectively, such as components, function calls and system calls.

Second, we designed THEA to build optimal classifier by handling three classes of features and then made a comprehensive judgment of unknown Android application. Finally, we implemented an automated tool named Androdect to detect 1126 malicious and 2000 non-malicious apps in real.

The experimental results show that Androdect plays the role of multi-class Android features in unknown malware detection and it performs better than other related works on the availability, efficiency and accuracy.

**Keywords** ensemble algorithm; Android application; multi-class features; malicious code detection;

Behavior analysis; data mining; smartphone; network behavior

收稿日期：2013-04-13；最终修改稿收到日期：2013-11-19.本课题得到国家自然科学基金（61272481）、北京市自然科学基金（4122089）、国家发改委信息安全专项（发改办高技[2012]1424号）资助.**杨 欢**，女，1984年生，博士，助理工程师，主要研究方向为信息安全、漏洞挖掘和恶意代码检测.E-mail:yangh@nipc.org.cn. **张玉清**，男，1966年生，博士，教授，中国计算机学会（CCF）会员，主要研究领域为网络与信息系统安全. **胡予濮，**男，1955年生，博士，教授，主要研究领域为序列密码与分组密码、网络安全协议与分析. **刘奇旭**，男，1984年生，博士，主要研究方向为信息安全、漏洞挖掘和漏洞等级评估.

1. 引言

近年来，智能手机普及率迅速增加，智能手机应用功能激增，全球最具权威的IT研究与顾问咨询公司Gartner的数据显示，预期到2016年年底，将有23亿部电脑、平板电脑以及只能手机使用Android系统，而微软的Windows设备的数量为22.8亿。Android几年来的快速发展时期成为了一个占主导地位的只能手机平台，拥有整个市场约三分之二的份额。由于Android系统的开放性，他也成为了众多恶意代码开发者的活跃地盘。

Android用户可以从Google Play和第三方Android市场（例如：Amazon、hiapk、gfan等）下载应用程序。应用程序开发者可以上传自己的应用程序到第三方Android市场。Google Play（前名为Android Market），是一个由Google 为Android设备开发的在线应用程序商店，可以让用户浏览、下载及购买Google Play上的第三方应用程序。Google Play提供了Google’s Bouncer检测恶意程序，但是实时性不够，恶意程序在检测出来之前已经被用户大量下载。

根据独立安全机构Av-test在2013年1月使用ESET、Lookout：Antivirus& security、Kaspersky、QIhoo360等22个手机杀毒软件对869个已知Android恶意应用的检测报告显示，各大手机杀毒软件的平均查杀率为94%。而对于未知恶意应用的检测，各大厂商没有公开杀毒软件的技术细节，本文无法做详细分析。由于恶意代码的数量和种类越来越多，加上代码混淆、加密等技术的兴起，使得恶意代码检测变得越来越困难。

目前，典型的恶意代码检测技术包括于签名、行为的检测方法。传统的基于签名的检测技术被普遍使用，但是他必须拥有一类恶意应用的签名库后才能检测该类恶意应用，因此无法有效检测未知的恶意应用。

基于行为的恶意代码检测目前主要采用动态和静态两种方法。动态方法是在系统运行过程中收集应用程序的一些行为信息。有点事绕过了静态方法遇到的代码混淆和加密等发面的问题，缺点是动态测试代码覆盖率低，并且有些恶意程序可以防止自身在模拟器下运行，挡在模拟器下运行时会自动崩溃；静态方法主要研究使用反汇编反编译技术或者在smali中间代码上进行控制流和数据流分析技术来进行恶意代码检测．优点是代码覆盖率高，缺点是 要解决静态方法无法检测的代码混淆、加密以及在 动态执行中才解码恶意代码的问题。

由于数据挖掘技术可以从大量数据中挖掘出有 意义的信息，有些研究者利用该技术挖掘出恶意 应用的行为特征，以此来检测未知恶意应用．但是， 目前的研究方法存在3方面的局限性：（1）不同的 数据挖掘算法针对同一类特征的检测效果不同，无法预知哪个算法效果最优；（2）同一种算法对不同 类型的特征检测效果不一定都是最优；（3）使用单 一的算法不能充分发挥每类特征在Android恶意应 用检测时所起的不同作用。

本文针对以上方法的局限性，首次提出一种基 于三层混合系综算法的多类特征Android应用恶意 行为检测系统，自动检测Android平台的恶意应用 程序．该系统采用动静态结合的行为特征提取方法， 提取Android应用的多类行为特征，再采用三层混 合系综算法建立检测模型，并对大量现实中的应用 程序进行检测．具体方法介绍如下： 首先，本文创建恶意应用库和非恶意应用库．恶 意应用库的创建参考文献[5]中所使用的部分恶意 应用．非恶意应用库的创建是通过修改Akdeniz的爬虫程序从Google play上批量下载应用程序， 并使用多种已有检测软件、杀毒软件和人工进行检 测，确保数据库的准确性． 其次，在特征提取阶段，本文采用动静态结合的 分析技术．静态分析主要对AndroidManifest.xml文件进行自动化分析．一是提packet名称和activity 名称为动态分析提供参数；二是对permission、activity、servicer和provider组件进行特征 提取；三是对使用native代码的应用程序中的lib库文件进行分析，提取应用程序的函数调用序列．动态分析主要采用沙盒技术，自动完成启动模拟器、安 装应用程序、模拟用户行为和提取系统调用序列等工作。

而后，在特征描述阶段，对提取的多类特征进行格式化处理，构建各类特征的集合．一是静态获取的 各个组件使用个数的统计数据；二是静态获取的应 用程序ｌｉｂ文件所使用的重要函数调用的统计数 据；三是动态获取的各个系统调用使用次数的统计 数据．

最后，使用三层混合系综算法建立检测模型．第 一层使用多个基础分类器对３种不同类型特征进行 训练，给出适 合不同类型特征的最优算法；第二层将 系综算法与最优分类算法结合再次进行训练建立模 型，进一步提高检测准确率； 第三层对待检测的应用 程序使用建立的模型进行分类并使用判决算法给出 检测结果．本方法通过实验对1126个恶意应用和2000个 非恶意应用进行了检测．实验结果表明，本文提出的 方法在有效性、准确性和执行效率上表现良好，准确率 为94.24%，优于其它Android恶

意应用检测系统。

本文的主要创新点与贡献如下：

（1）采用动静态结合的行为特征提取方法，动态方法弥补了静态方法不能检测运行中释放的恶意 代码的缺点，也绕过了静态方法遇到的代码加密和 混淆问题；静态方法为动态方法提供了运行参数，可以分析动态方法无法处理的应用程序。

（2）本文设计的检测系统提取了应用程序的多 类行为特征，充分反映了应用程序行为，并且该系统 具有可扩展性，可以加入更多类型特征进行检测．

（3）首次提出了一种三层混合系综算法THEA（Triple Hybrid Ensemble Algorithm），对各类特征 分别选取最优的基础分类算法从而给出综合判决结 果.并首次将其应用到Android应用的恶意代码检测，实验证明了该算法的有效性．

（4）将提出的原型系统实现了自动化的检测工具Androdect，对大量现实中的恶意应用和非恶意应用进行检测.

本文第2节介绍相关工作；第3节总体介绍检 测系统；第4节阐述系统的关键技术和算法；第5节 给出实验设计与结果分析；第6节讨论本文提出系 统的局限性和下一步工作；最后在第7节对本文进 行总结

感悟：通过这次作业，我掌握了Word的一些文本编辑的基本知识（页眉，排版，模板，字体，段落等），并初步掌握了论文书写的格式，为以后的可能会有的论文写作打下了基础。