

联合选择特征和分类器参数的 Android 系统安全检测

陈泽恩

(惠州学院 教育技术中心, 广东 惠州 51600)

摘 要: 为了提高 Android 系统安全检测性能, 提出一种联合选择特征和分类器参数的 Android 系统安全检测模型. 首先提取 Android 系统安全检测特征, 并将特征和支持向量机参数组合在一起作为一个上体, 然后通过教与学优化算法模拟老师的教学过程和同学之间的互相交流过程, 找到最优特征子集和支持向量机参数, 最后构建最优的 Android 系统安全检测模型, 并进行仿真实验. 实验结果表明, 相对于其他 Android 系统安全检测模型, 模型提高了 Android 系统安全检测准确率, 改善了 Android 系统安全检测效率, 可以满足 Android 系统安全检测的实时性要求.

关键词: 安全检测; 特征选择; 分类器设计; 教与学优化算法

中图分类号: TP393

文献标识码: A

文章编号: 1000-7180(2015)02-0119-05

Secure Detection of Android System by Jointly Selecting Features and Classifier Design

CHEN Ze-en

(Center of Educational Technology, Huizhou University, Huizhou 516007, China)

Abstract: In order to improve the performance of android security detection, this paper puts forward a secure detection model of android applications by jointly selecting features and classifier design. Firstly, android security detection features are extracted, and the feature and support vector machine parameter are combined to a Individual, and then teaching learning optimization algorithm is used to simulate teacher's teaching process and students exchange process to find the best subset of features and support vector machine parameters, finally, the optimal android security detection model are established and the simulation experiment is carried. The experimental results show that, compared with other android security detection model, the proposed model has improved the accuracy rate of android security detection, and improved android security detection efficiency, so it can meet the real-time requirement of android security detection.

Key words: secure detection analog circuits; features selecting; classifier design; teaching learning optimization algorithm

1 引言

随着超大规模集成电路技术的不断成熟, 智能手机和平板电脑成为人们生活的必需品, Android 系统具有开放、运行速度快等优点, 在智能终端中得到了广泛的应用^[1]. 由于 Android 系统的开放性, 基于 Android 的恶意软件猛增, 用户切身利益受到了

威胁, 为此如何对恶意软件进行检测, 保证 Android 系统应用过程中的安全性, 成为了 Android 技术中的研究重点和难点^[2].

针对 Android 的安全性能检测问题, 广大学者进行了深入的研究, 传统 Android 安全性能检测方法主要基于签名技术, 该方法必须对恶意软件具有一定的先验知识, 而且不能检测未知或者变种的恶

意软件,因此在实际应用中受到一定的限制^[3].当前一般采用基于行为的 Android 安全性能检测方法,其分为动态和静态两种检测技术^[4].动态技术首先收集应用程序的一些行为,不需要静态技术的加密算法,通过文件权限改变、线程运行情况进行恶意软件检测,但是存在采集应用程序的动态数据时间长,检测效率低等不足^[5].静态技术具有简单、快速等优点,主要包括反编译、模式匹配等分析技术,但是其存在误报率、漏报率比较高等不足^[6].由于动态和静态检测技术均存在各自的缺陷,目前一般的 Android 安全性能方法均同时包括动态和静态检测技术.Android 安全性能检测实际上是一种分类问题,其包括特征选择和分类器设计^[7].为了获得更好的 Android 安全性能检测效果,通常情况下,需要尽可能多地提取特征,这样特征中会存在许多冗余特征以及对检测结果无用特征,影响了 Android 安全性能检测正确率和效率.当前特征选择算法主要有遗传算法、粒子群算法等,但这些方法均存在早熟和误收敛的问题^[8].Android 系统安全检测分类器主要采用神经网络、隐马尔可夫模型以及支持向量机等,其中神经网络要求样本多,检测成本高;隐马尔可夫模型具有严谨的数据结构和可靠的计算性能,但是没有自学习能力;支持向量机具有分类能力强,泛化能力好等优点,为此本文利用支持向量机设计 Android 系统安全检测的分类器^[9-12].

教与学优化 (Teaching Learning Optimization, TLO) 算法模拟教师与学生之间交流过程的智能算法,具有鲁棒性好、易实现等优点^[11],为了提高 Android 系统安全检测性能,提出一种联合选择特征和分类器参数的 Android 系统安全检测模型,采用 TLO 算法对 Android 安全特征和分类器参数进行联合优化,最后通过仿真实验对其性能进行验证.

2 Android 安全检测的特征选择和分类器设计

2.1 安全检测的特征选择数学模型

首先分别提取 Android 安全性能检测的组件特征、函数调用特征以及系统调用特征,它们组成一个特征集 $S = \{s_1, s_2, \dots, s_n\}$, n 表示是特征个数,Android 安全性能特征选择主要是检测正确率(G),那么 Android 安全性能检测的特征选择数学模型可以采用式(1)进行描述:

$$\begin{aligned} & \max_S G(S) \\ & s. t. \\ & \begin{cases} S = \{s_1, s_2, \dots, s_n\} \\ s_i \in \{0, 1\} \\ i = 1, 2, \dots, n \end{cases} \end{aligned} \quad (1)$$

式中, $s_i \in \{0, 1\}$, 1 表示该特征选择, 0 分别表示该特征不在特征子集中.

2.2 分类器参数优化的数学模型

设 Android 系统安全检测的训练样本集为: $(x_i, y_i), i = 1, 2, \dots, n$, x_i 表示输入特征向量; y_i 是输出, 支持向量机分类平面的方程可以表达为

$$w \cdot x + b = 0 \quad (2)$$

基于结构风险最小化原则, 式(2)可转化成一个带约束的最小值的问题:

$$\begin{aligned} & \min \frac{1}{2} ||w||^2 \\ & s. t. \\ & y_i[(w \cdot x_i) + b] - 1 \geq 0 \\ & i = 1, 2, \dots, n \end{aligned} \quad (3)$$

引入松弛变量 $\xi = \{\xi_1, \xi_2, \dots, \xi_n\}$, 式(3)变成如下优化问题:

$$\begin{aligned} & \min \frac{1}{2} ||w||^2 + C \sum_{i=1}^n \xi_i \\ & s. t. \\ & y_i[(w \cdot x_i) + b] \geq 1 - \xi_i \quad (i = 1, 2, \dots, n) \\ & \xi_i \geq 0 \end{aligned} \quad (4)$$

式中, C 为惩罚参数.

选择高斯径向基函数作为支持向量机的核函数, 那么支持向量机的最优分类函数为

$$\begin{aligned} f(x) &= \text{sgn}\{(w^* \cdot x_i) + b^*\} \\ &= \text{sgn}\left\{\left(\sum_{j=1}^n y_j a_j^* K(x_i, x_j)\right) + b\right\} \\ &= \text{sgn}\left\{\left(\sum_{j=1}^n y_j a_j^* \exp\left(-\frac{|x_i - x_j|^2}{\sigma^2}\right)\right) + b\right\} \end{aligned} \quad (5)$$

采用支持向量机设计 Android 安全性能检测的分类器时, 其参数 σ 和 C 相关, 支持向量机参数优化的目标为 Android 系统安全检测正确率(G), 其数学模型为

$$\begin{aligned} & \max_M G(C, \sigma) \\ & s. t. \\ & \begin{cases} C \in (C_{\max}, C_{\min}) \\ \sigma \in (\sigma_{\min}, \sigma_{\max}) \end{cases} \end{aligned} \quad (6)$$

2.3 联合特征选择和分类器参数的数学模型

在 Android 系统安全检测中,由于检测性能同时与特征选择和分类器设计密切相关,而且当前算法均没有考虑两者之间联系,分开优化,难以获得最理想的 Android 系统安全检测结果,为此,本文综合特征选择和分类器参数之间的联系,对它们进行联合优化,联合选择特征和分类器参数的数学模型可以采用式(7)进行描述:

$$\begin{aligned} & \max \{G(S, C, \sigma)\} \\ & s. t. \\ & \begin{cases} C \in (C_{\min}, C_{\max}) \\ \sigma \in (\sigma_{\min}, \sigma_{\max}) \\ S = \{s_1, s_2, \dots, s_n\} \end{cases} \end{aligned} \quad (7)$$

3 联合选择特征和分类器参数的 Android 安全检测模型

3.1 教与学优化算法

TLO 算法是受老师教学过程和学生学习过程的后发而提出来的,与粒子群算法、遗传算法相比,有效地利用了群体均值在搜索过程中的指导作用,同时对不同个体进行差异化学习,以获得较高的搜索性能。TLO 算法的工作步骤如下:

(1) 种群初始化。采用随机初始化的方法,产生 NP 个 D 维变量作为学生群体,计算每个学生的适应度值,找出其中的最优个体作为老师。

$$x_{i,j} = x_{j,\min} + \text{rand} \times (x_{j,\max} - x_{j,\min}) \quad (8)$$

式中, NP 为学生的数量; $x_{\min}$ 和 $x_{\max}$ 分别为变量值的上下限, rand 为 0 到 1 之间的随机数^[12]。

(2) 教学阶段。模拟老师的教学过程,通过缩小老师与学生平均成绩之间的差距,提高班级的整体水平。教学过程的具体表达式如下:

$$X^{\text{new}} = X^{\text{old}} + \text{rand} \times (X^{\text{teacher}} - \text{TF} \times \text{Mean}) \quad (9)$$

式中, TF 是学习因子, Mean 是学生各科成绩的平均值, X^{old} 和 X^{new} 为教学过程前后学生的成绩。

(3) 学习阶段。学生群体经过统一的教学过程后,个体之间通过互相交流进行差异化学习。从学生群体中随机选择两个个体 X^{r1} 和 X^{r2} , 比较两者适应度值的大小,选择较优的个体进行学习。对于最小值问题,差异化学习的具体表达式如式(10)所示。

$$x^{\text{new}} = \begin{cases} X^{\text{old}} + \text{rand} \cdot (X^{r1} - X^{\text{old}}), & f(X^{r1}) < f(X^{r2}) \\ X^{\text{old}} + \text{rand} \cdot (X^{r2} - X^{\text{old}}), & f(X^{r1}) \geq f(X^{r2}) \end{cases} \quad (10)$$

式中, X^{r1} 和 X^{r2} 为两个不同个体。

(4) 终止判断。判断是否满足终止条件,若达到最大迭代次数,终止迭代,否则,返回步骤(2),继续执行教学过程和学习过程。

3.2 个体的编码方式设计

TLO 算法的个体编码方式是所求解问题进行对接,由于本文目标不是找到最优特征和最优分类器,因此个体包括三部分:特征子集,支持向量机的惩罚参数 C, 高斯径向基函数参数 σ , 采用二进制方式对它们进行编码,具体如图 1 所示。

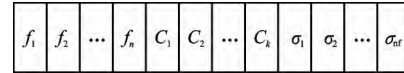


图 1 个体的编码方式

3.3 适应度函数设计

由于特征选择和分类器设计主要获得更加理想的 Android 系统安全检测结果,因此采用 Android 系统安全检测正确率作为适应度函数

3.4 Android 系统安全检测模型的工作步骤

(1) 收集 Android 系统安全检测的过程特征信息,为了避免因特征值差异对检测结果的干扰,采用式(11)对特征值进行预处理:

$$x_i' = \frac{x_i}{\max(x_i)} \quad (11)$$

式中, $\max(x_i)$ 为特征 x_i 的最大值。

(2) 设置支持向量机参数的范围以 TLO 算法的参数。

(3) 初始化种群。随机产生 m 个学生个体,每一个个体由特征子集,支持向量机的惩罚参数 C, 高斯径向基函数参数 σ 组成。

(4) 对个体进行解码,采用特征子集对训练样本的特征进行选择,并输入到支持向量机学习,支持向量机参数为解码的 C 和 σ , 根据学习结果计算个体的适应度值,找到其中的最优个体作为老师。

(5) 根据式(9)向老师学习,缩小班级平均水平与老师之间的差距,从而提高班级的整体水平。

(6) 在老师的教学阶段后,个体之间通过式(3)进行自由的差异化交流,通过向邻近较优的个体学习,来提高自己的适应度值。

(7) 当达到最大迭代次数,则停止优化,不然返回步骤(5)继续执行迭代学习。

(8) 根据最优特征子集和最优支持向量机参数建立 Android 系统安全检测模型。

综合上述可知,基于联合特征选择和分类器设计的 Android 系统安全检测模型工作流程如图 2 所示。

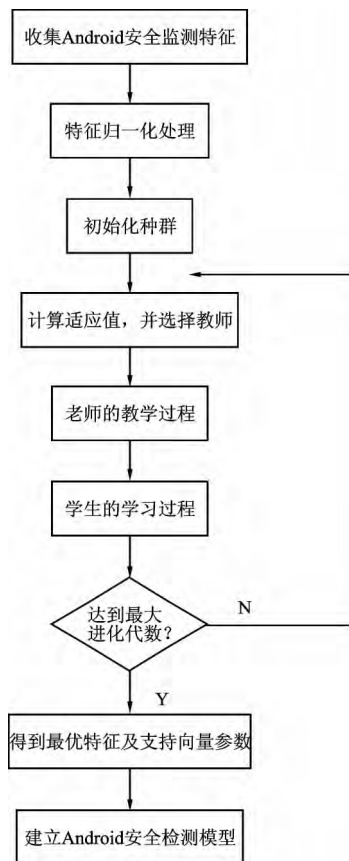


图2 Android 系统安全检测模型的工作流程

4 仿真实验

4.1 仿真环境

为了测试本文提出 Android 系统安全检测模型的性能,实验数据中恶意样本集来源于北卡罗来纳州立大学的恶意样本数据库,正常样本来源于 Google Play 商店随机下载.为了使本文模型的结果具有可比性,采用表 1 所示的模型进行对比实验.采用 Android 系统安全检测正确率、误报率以及平均检测时间(秒,s)对它们检测结果进行分析.

表 1 仿真实验的 Android 系统安全检测对比模型

模型名	模型描述
SVM1	全部特征+随机确定支持向量机参数
SVM2	TLO 算法选择特征+随机确定支持向量机参数
SVM3	全部特征+TLO 算法选择支持向量机参数
SVM4	TLO 算法选择特征+TLO 算法选择支持向量机参数,但两者分开进行

4.2 结果与分析

4.2.1 检测正确率和误报率对比

采用 SVM1、SVM2、SVM3、SVM4 以及本文模型对收集的样本数据进行学习,建立 Android 系统安全检测模型,每一种模型均运行 5 次,它们的检测

正确率和误报率分别如图 3 和图 4 所示.对图 3 和图 4 的结果进行分析,而且得到如下结论.

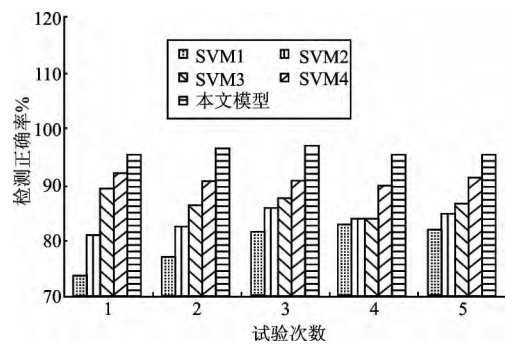


图 3 不同 Android 系统安全检测模型的正确率对比

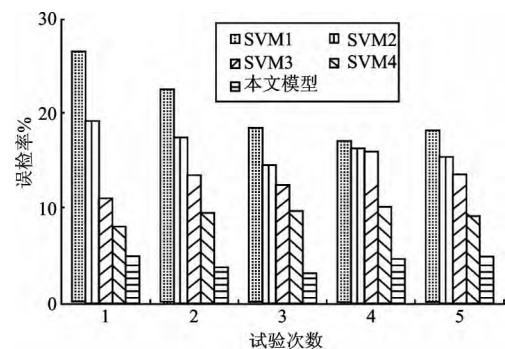


图 4 不同 Android 系统安全检测模型的误报率对比

(1)相对于基于 SVM1 的 Android 系统安全检测模型,基于 SVM2 的 Android 系统安全检测模型的性能得到了改善,提高了检测正确率,降低了误报率,这表明,收集的 Android 系统安全检测特征中存在一些对检测结果无用或者有负面影响特征,通过 TLO 算法对特征进行选择,可以提高 Android 系统安全检测性能.

(2)相对于基于 SVM1 的 Android 系统安全检测模型,基于 SVM3 的 Android 系统安全检测结果更优,这主要是由于支持向量机的分类性能与其参数密切相关,采用人工方式设置参数,建立的 Android 系统安全检测模型的检测正确率比较低,检测结果不可靠,而通过 TLO 算法对支持向量机参数进行优化,可以设计更优的 Android 系统安全检测分类器,能够进行最优特征选择,有效降低了 Android 系统安全检测的误报率.

(3)相对于对比的 Android 系统安全检测模型,本文模型的 Android 系统安全检测准确率大幅度提高,而且检测结果更加稳定,这主要由于单独对特征进行选择,或者分开实施分类器以及特征选择均难以建立整体性能更优的 Android 系统安全检测模型,主要是它们割裂特征和分类器对检测结果的影

响,而本文采用 TLO 算法对支持向量机参数和特征进行联合选择和优化,不仅提取了检测结果贡献大的特征,而且找到了最优的支持向量参数,提高了 Android 系统安全检测效果,具有明显的优势.

4.2.2 检测效率对比

在 Android 系统安全检测实际应用中,恶意软件检测的效率相当重要,所有模型平均检测时间如表 2 所示.从表 2 可以清楚看出,相对于其他的 Android 系统安全检测模型,本文提出的 Android 系统安全检测模型的平均检测时间最短,检测效率更高,这主要是由于通过 TLO 算法对特征进行选择,大幅度减少了支持向量机的输入维数,降低了计算复杂度,同时采用 TLO 算法对参数优化,提高了支持向量机的学习速度,可以满足 Android 系统安全检测应用过程中的实时根据.

表 2 不同 Android 系统安全检测模型的效率对比

Android 系统安全检测模型	平均检测时间/s
SVM1	4.11
SVM2	2.53
SVM3	3.93
SVM4	3.73
本文模型	2.51

5 结束语

针对当前 Android 系统安全检测正确率的难题,在分析特征选择与分类器设计之间联系的基础上,提出一种联合选择特征和分类器参数的 Android 系统安全检测模型.仿真结果表明,相对于其他的 Android 系统安全检测模型,本文模型通过 TLO 算法选择最优的特征子集,减少了特征维数,提高了 Android 系统安全检测效率,同时通过 TLO 算法选择分类器参数,提高了 Android 系统安全检测正确率,可以减少恶意软件对 Android 系统的威胁,具有较高的实际应用价值.

参考文献:

[1] Zhou Ya Jin, Jiang Xu Xian. Dissecting android malware; characterization and evolution [C]// IEEE Symposium on Security and Privacy. San Francisco, CA: IEEE,2012: 95-109.

[2] Shabtai A, Kanonov U, Elovci Y, et al. Abnormally;

a behavioral malware detection framework for android devices [J]. Journal of Intelligent Information Systems,2012, 38(1): 161-90.

[3] Burguera I, Zurutuza U, Nadjmehrani S. Crowdroid: behavior based malware detection system for android [C]// Proceedings of the ACM Workshop on Security and Privacy in Smart phones and Mobile Devices. New York:ACM,2011:15-6.

[4] Dini G, Martinelli F, Saracino A, et al. MADAM: a multi-level anomaly detector for android malware [M]. Heidelberg: Springer Berlin ComputerNetwork Security, 2012: 240-53.

[5] 李静华,慕德俊,杨鸣坤,等. Android 恶意程序行为分析系统设计 [J]. 北京邮电大学学报, 2014 (27): 104-107.

[6] 文伟平,梅瑞,宁戈,等. Android 恶意软件检测技术分析和应用研究[J]. 通信学报, 2014,35(8):78-74.

[7] Shabtai A, Moskovitch R, Feher C, et al. Detecting unknown malicious code by applying classification techniques on Opcode patterns [J]. Security Informatics, 2012, 1(1):1-22.

[8] Bente I, Hellmann B, Vieweg J, et al. TCADS: trustworthy context-related anomaly detection for smart phones [C]// International Conference on Network-Based Information Systems Nelbour. Australia:IEEE, 2012: 247 254.

[9] 张玉清,胡予蹊,刘奇旭. 基于多类特征的 Android 应用恶意行为检测系统[J]. 计算机学报, 2014, 37(1): 15-27.

[10] Padriya Nitin, Mistry Nilay. Review of behavior malware analysis for Android [J]. International Journal of Engineering and Innovative Technology (IJEIT), 2013,2(7):230-232.

[11] Rao R V, Savsani V J, Vakharia D P. Teaching-learning-based optimization: a novel method for constrained mechanical design optimization problems [J]. Computer Aided Design, 2011, 43(3): 303-315.

[12] Rao R V, Savsani V J, Vakharia D P. Teaching-learning-based optimization: An optimization method for continuous non-linear large scale problems [J]. Information Sciences, 2012, 183 (1): 1-15.

作者简介:

陈泽恩 男,(1976-),硕士,实验师.研究方向为计算机软件技术、教育技术、实验室管理.