◎网络、通信、安全◎

面向Android手机平台异常入侵检测的研究

杨午圣,孙 敏
YANG Wusheng, SUN Min

山西大学 计算机与信息技术学院,太原 030006
School of Computer and Information Technology, Shanxi University, Taiyuan 030006, China

YANG Wusheng, SUN Min. Research on Android mobile phone platform for anomaly intrusion detection. Computer Engineering and Applications, 2014, 50(7):71-74.

Abstract: With the popularity of smart mobile phone, the harm of intrusion is more and more serious. This paper is, based on the Android smartphone platform, combined with intrusion detection research, to solve the problem of intrusion detection of smartphone. In order to give a reasonable judgment and update the phone as soon as possible, the paper collects the system and the network characteristic data on the Android platform, and uploads them to the remote cloud servers, then analyzes using Support Vector Machine(SVM). The experimental results show that, taking the kind of mechanism not only can reduce resource consumption of smart phones, but also can handle and response the intrusion as quickly as possible.
Key words: smartphone; Android platform; intrusion detection; Support Vecior Machina(SVM)

摘 要: 智能手机应用普及的同时,入侵的危害也越来越严重。针对Android智能手机平台,结合入侵检测的相关研究,解决智能手机入侵检测的问题。采取在Android平台下采集系统和网络特征数据,上传至远程云服务器,在服务器上利用SVM进行分析处理,以给出合理的入侵与否的判断,进而尽快更新手机的处理机制。实验结果表明,既减少了智能手机资源消耗,又能对手机的异常入侵尽快做出反应和处理。
关键词: 智能手机; Android平台; 入侵检测; 支持向量机(SVM)
文献标志码: A **中图分类号:** TP309 doi:10.3778/j.issn.1002-8331.1309-0056

1 引言

随着3G通信网络的普及,4G网络的兴起,智能手机尤以Android智能手机占市场份额最大。

由于Android用户数庞大,开源性强,用户可自行安装软件、游戏等第三方提供的程序,而这些信息对于许多用户而言并不知道是否安全,很多异常入侵攻击者把矛头指向了它[1]。不仅如此,随着手机智能技术的发展,针对智能手机的异常入侵也跟着变得多样起来。虽然Android平台的开源、开放、免费等特性为Google带来了大量的市场占有率,但是这也给消费者带来了不少安全隐患。每个人手机中的个人隐私,一旦外泄,给用户带来的损失是无法估计的,那么智能手机主动防御病毒入侵的研究迫在眉睫。

2 相关研究

早在1980年,入侵检测相关研究就已经开始[2],但是在智能手机入侵检测方面,可以说才刚开始起步。

在移动通信网络中,第一个IDS是为AMPS模拟蜂窝系统设计的[3]。

而在智能手机的异常检测研究中,傅德胜等[4]提出的手机轻量型入侵检测系统模型,将Snort技术应用于手机领域。但它一般是用于检测网络类型的入侵检测。

在2010年,Shabtai等提出了基于知识的时间序列的方法[5],这种方法能有效利用智能手机的时间序列的特征,也比较符合智能手机的入侵检测系统.

由于智能手机的入侵检测研究并不多,以上对智能手机的研究[6],取得了较好的成果。在智能手机领域中,

基金项目: 国家自然科学基金(No.61100058);山西省自然科学基金(No.2011011014-2);山西省高校教学改革重点项目(No.J2013010)。
作者简介: 杨午圣(1989—),男,硕士,主要研究领域为计算机网络安全;孙敏(1966—),通讯作者,女,副教授。E-mail:minsun@sxu.edu.cn
收稿日期: 2013-09-05 **修回日期:** 2013-11-29 **文章编号:** 1002-8331(2014)07-0071-04

异常入侵检测也在不断的探索过程中。

3　面向Android平台入侵检测模型

本文检测模型总体框架[7],如下图1所示。

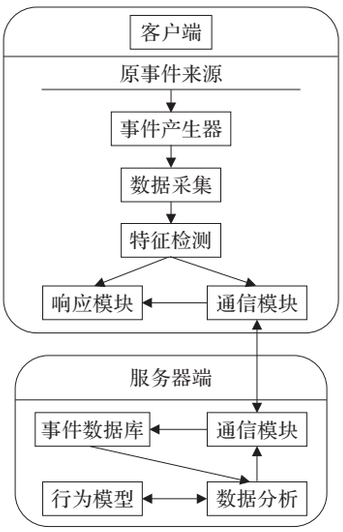

图1　模型总体框架图

3.1　Android客户端模型设计

该客户端主要用来收集最原始的数据,即事件来源[8],同时,做出一些简单的操作判断,并能对服务器端返回的判断结果做出相应的反应,让用户选择相应的操作。

客户端分为6大数据采集模块,分别为CPU信息采集模块、内存信息采集模块、网络信息采集模块、进程信息采集模块、磁盘信息采集模块以及短信采集模块。客户端将事件来源的数据进行初步整理,初步整理后的数据传给数据通信模块;数据通信模块是与云服务器连接的桥梁,它将初步整理的数据上传至模拟云服务器,同时又随时接收模拟云服务器返回的消息,并传给客户端的数据响应模块;数据响应模块根据数据来产生响应,如切断网络,禁止访问,删除文件,发出警报,弹出提示等。

3.2　服务器端模型设计

服务器端同样也有通信模块,能过通信模块接收来自客户端传来的数据,进行简单过滤;将过滤后的数据放至事件数据库,数据处理模块从事件数据库中取出事件,利用事件分析模块进行分析;将分析后的结果再通过通信模块返回给客户端,客户端又利用通信模块接收服务器端的返回数据。

3.3　模型特征选取和实验算法

本文选Android作为实验平台[9],实验时先利用客户端模拟程序[10],在正常情况下,利用客户端的数据采集模块提取5天或1周时间的主体活动的基本信息;然后用程序模拟异常攻击,并采集相应数据。本文模拟了4种常见的异常:短时间向外发送大量手机短信;短时间内不停的接收手机短信;在用户打开发数据流量的情况下,不停地浪费用户的手机流量;在一段时间内,向手机本地存储卡内写入恶意文件。

3.3.1　本文选取的特征属性

根据要研究和模拟的异常,本文选取了17个关键特征属性,涵盖了研究所需的特征属性。分别为CPU信息:cpu_usage;内存信息:mem_usage、mem_cached、mem_active、mem_inactive、mem_active(anon)、mem_inactive(anon)、mem_active(file)、mem_inactive(file);网络信息:int_output、int_input、int_tcp、int_udp;磁盘信息:/SD_card;进程信息:process_number;短信信息:message_send、message_received。上述都为连续属性。

3.3.2　支持向量机

支持向量机是Cortes和Vapnik[11]于1995年首先提出的,它是建立在统计学习理论的VC维理论和结构风险最小原理基础上的,根据有限的样本信息在模型的复杂性和学习能力之间寻求最佳折衷,以期获得最好的推广能力(或称泛化能力)。

基于支持向量机的特点和优势,本文采用*C*-SVC和*v*-SVC两种类型SVM来进行实验。实验证明了支持向量机很适合本文所研究的内容。以下为两种算法的简单描述。

(1)*C*-SVC算法

C-SVC算法[12-13]通常是标准的分类算法,也主要是针对两类分类问题的支持向量机算法。它通过选取核函数替代变换,用核函数代替内积来使问题变得更简单计算。算法思想:

设样本 $\boldsymbol{x}$ 为 n 维向量,即 $\boldsymbol{x}_i \in R^n$,其中 $i=1,2,\cdots,k$,而 $\boldsymbol{y} \in R^k$,$\boldsymbol{y} \in \{1,-1)\}$。由线性支持向机可知:*C*-SVC算法解决的是下面的原始问题:

$$\min_{\omega,b,\varepsilon} \frac{1}{2}\boldsymbol{W}^{\mathrm{T}}\boldsymbol{W}+C\sum_{i=1}^{k}\varepsilon_i$$

$$\text{s.t.}\quad y_i((\boldsymbol{W}^{\mathrm{T}}\cdot\varnothing(\boldsymbol{x}_i))+b)\geqslant 1-\varepsilon_i$$

其中 $\varepsilon_i \geqslant 0$,且 $i=1,2,\cdots,k$。

然后用 a 替代 $\boldsymbol{W}$ 得到其对偶问题,再根据核函数变换 $K(\boldsymbol{x}_i,\boldsymbol{x}_j)=\varnothing(\boldsymbol{x}_i)^{\mathrm{T}}\varnothing(\boldsymbol{x}_j)$,将训练集 $\boldsymbol{x}_i$ 映射到高维空间,最终决策函数变成:

$$\mathrm{sgn}(\sum_{i=1}^{k} y_i a_i K(\boldsymbol{x}_i,\boldsymbol{x}_j)+b)$$

(2)*v*-SVC算法

在前面的*C*-SVC算法中,惩罚因子 C 本身没并没有什么确切的意义。而*v*-SVC[13-14]算法与*C*-SVC算法不同的是,它利用一些具有意义的参数 v 来取代了 C。算法思想:

设样本 $\boldsymbol{x}$ 为 n 维向量,即 $\boldsymbol{x}_i \in R^n$,其中 $i=1,2,\cdots,k$,而 $\boldsymbol{y} \in R^k$,$\boldsymbol{y} \in \{1,-1)\}$。则其原始问题变成了:

$$\min_{\omega,b,\varepsilon,\rho}\frac{1}{2}\boldsymbol{W}^{\mathrm{T}}\boldsymbol{W}-v\rho+\frac{1}{k}\sum_{i=1}^{k}\varepsilon_i$$

$$\text{s.t.}\quad y_i((\boldsymbol{W}^{\mathrm{T}}\cdot\varnothing(\boldsymbol{x}_i))+b)\geqslant\rho-\varepsilon_i$$

其中 $\varepsilon_i\geqslant 0$,且 $i=1,2,\cdots,k$,且 $\rho\geqslant 0$ 。

然后同样得出其对偶问题,最后可得到决策函数为:

$$\operatorname{sgn}(\sum_{i=1}^{k}\frac{y_i a_i K(\boldsymbol{x}_i,\boldsymbol{x}_j)}{\rho}+b)$$

4　实验结果与分析

4.1　实验数据描述

本文用模拟器提取了5天时间主体活动的基本信息,经过整理处理后,得到了相对完备的数据集。选用了1 257条数据作为训练数据,其中异常样本255条,约占20%;而测试时,共取859条数据作为测试数据,考虑到实际情况,取异常样本89条,约占10%,将其混合到测试样本中进行实验。

由于每个属性值范围不同,本文对所有数据进行了规范化,将其数值属性规范化到[0,1]中。公式为:

$$Value=lower+(upper-lower)\times(value-\min(f_i))/(\max(f_i)-\min(f_i))$$

其中 *upper* 表示规范后的上界,这里为1; *lower* 为规范后的下界,这里为0; $\max(f_i)$、$\min(f_i)$ 分别表示属性 f_i 的最大值以及最小值。

实验结果采用正确率(*CR*)、检测率(*DR*)和误报率(*ER*)作为实验的评估标准[15]:

正确率=(正确分类样本数/总的样本数)×100%

检测率=(正确检测出的入侵数/总的入侵样本数)×100%

误报率=(错误检测出的入侵数/总的样本数)×100%

4.2　不同参数选取

4.2.1　核函数的选择

实验比较了常见的4种核函数的结果,默认的惩罚因子 $C=1$ 。实验结果如表1~4所示。

表1　线性核函数　(%)

CR	*DR*	*ER*
97.67	92.14	1.51

表2　多项式核函数 ($\gamma=2$)　(%)

	D=2	*D*=3	*D*=4	*D*=5	*D*=6
CR	94.88	94.18	94.18	93.95	94.06
DR	89.89	86.52	88.76	91.01	92.14
ER	4.08	4.31	4.66	5.12	5.12

表3　RBF核函数　(%)

	$\gamma=1$	$\gamma=2$	$\gamma=3$	$\gamma=4$	$\gamma=5$
CR	96.74	96.62	96.74	96.86	97.32
DR	89.89	89.89	89.89	88.76	86.52
ER	2.21	2.33	2.21	1.98	1.28

表4　Sigmoid核函数　(%)

	$\gamma=0$	$\gamma=0.1$	$\gamma=0.2$	$\gamma=0.3$	$\gamma=0.4$
CR	97.44	97.67	98.02	97.79	97.56
DR	79.78	87.64	93.26	94.38	92.12
ER	0.47	1.05	1.28	1.63	1.63

从实验结果可以看出,所有核函数总体的检测率和正确率随着参数的变大呈现先增后减的趋势;而对误报率而言,随着检测率的变化而变化,当检测率提高时,误报率也跟着升高了。综合考虑,当 $C=1$ 时,对于本实验取Sigmoid函数,总体结果相对较好。

4.2.2　惩罚因子 *C* 的选择

惩罚因子 *C* 作为SVM的一个重要参数,影响着分类器的推广泛化能力。它决定了你有多重视离群点带来的损失,当所有离群点的松弛变量的和一定时,你定的 *C* 越大,对目标函数的损失也越大。而在高维特征向量空间上,边缘上的支撑向量会随着惩罚因子 *C* 的变化而变化。一般当 *C* 减小时,其偏离程度也随之减小;反之,则偏离程度加大。本文取 *C*=10,50,100,1 000,10 000,分别进行测试实验,实验结果如表5~7所示。

表5　多项式核函数 ($g=1/k, d=3$)　(%)

	C=10	*C*=50	*C*=100	*C*=1 000	*C*=10 000
CR	96.86	97.79	97.09	94.65	94.65
DR	74.16	89.89	88.76	86.52	84.27
ER	0.47	1.16	1.75	3.96	3.73

表6　RBF核函数　(%)

	C=10	*C*=50	*C*=100	*C*=1 000	*C*=10 000
CR	97.32	97.79	96.39	94.65	93.83
DR	92.14	89.89	89.89	86.52	78.65
ER	1.86	1.16	2.56	3.96	3.96

表7　Sigmoid核函数　(%)

	C=10	*C*=50	*C*=100	*C*=1 000	*C*=10 000
CR	97.55	97.32	97.32	97.44	96.51
DR	91.01	89.89	89.89	96.51	92.14
ER	1.51	1.63	1.63	1.86	2.68

当取不同的惩罚因子 *C* 时,分别比较其检测率以及误报率,得到结果如图2和图3所示。

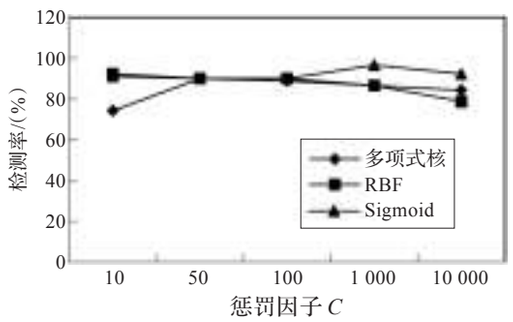

图2　检测率随 *C* 的变化

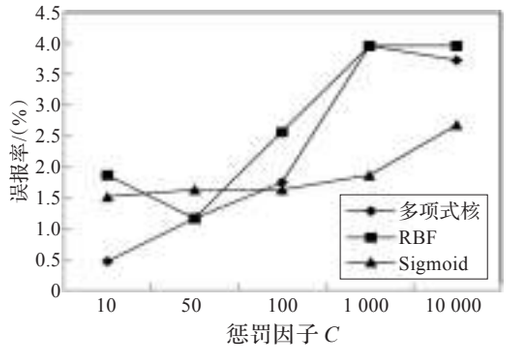

图3　误报率随 C 的变化

由以上实验结果可以看出,检测率除了RBF函数随着 C 的增大而减小,另外两种函数总体呈现先增后减的趋势。另外,误报率随着 C 的增大总体也呈现增大趋势。

4.2.3　参数 v 的选择

以上为选用 C-SVC 算法时的整个实验结果,下面是选用 v-SVC 算法时参数 v 对实验结果的影响。当 v 取值不当时,实验将无分类结果。本文取 $v=0.10$,0.15,0.20,0.25,0.30分别进行实验,得到结果如表8。

表8　各函数实验结果　(%)

核函数		v=0.10	v=0.15	v=0.20	v=0.25	v=0.30
线性	*CR*	97.32	97.56	97.67	98.02	98.02
	DR	89.89	92.14	89.89	93.26	88.76
	ER	1.63	1.63	1.28	1.28	0.81
多项式	*CR*	96.04	97.09	97.56	97.91	97.67
	DR	86.52	87.64	88.76	89.89	86.52
	ER	2.56	1.63	1.28	1.05	0.93
RBF	*CR*	96.28	97.09	97.56	97.91	97.91
	DR	88.76	89.89	91.01	92.14	88.76
	ER	2.56	1.86	1.51	1.28	0.93
Sigmoid	*CR*	97.21	97.79	97.67	97.91	97.91
	DR	89.89	93.26	89.89	93.26	88.76
	ER	1.75	1.51	1.28	1.40	0.93

随着 v 值的变化,检测率和误报率的变化如图4、5所示。实验结果表明,由于参数 v 的变化,检测率随着 v 的增大而先变大后变小,而误报率则随着 v 的增大总体呈现下降的趋势;其正确率总体也较高。

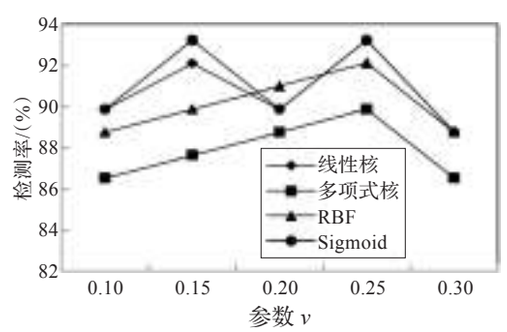

图4　检测率随 v 的变化

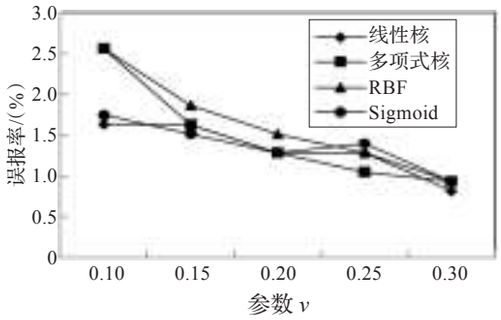

图5　误报率随 v 的变化

综合以上结果,本文根据正确率,同时参考其相应的检测率和误报率,选取最优 C 以及最优 v ,得到结果如表9、表10所示。

表9　最优 C 实验结果　(%)

	线性核	多项式核	RBF	Sigmoid
CR	97.67	97.79	97.32	97.55
DR	92.14	89.89	92.14	91.01
ER	1.51	1.16	1.86	1.51

表10　最优 v 实验结果　(%)

	线性核	多项式核	RBF	Sigmoid
CR	98.02	97.91	97.91	97.91
DR	93.26	89.89	92.14	93.26
ER	1.28	1.05	1.28	1.40

从以上结果可以看出,通过选择最优参数,可以得出较好的结果。改进后的 v-SVC 算法相对于 C-SVC 算法,实验结果还也所好转,其正确率以及检测率相对 C-SVC 而言有提高,而误报率则有所下降,这说明 v-SVC 算法更适合本实验所研究的问题。

5　结束语

对于智能手机异常入侵的研究,还需要进行大量深入的研究。同时,由于受到数据采集的限制,实验的完备性可能有所欠缺,这会对实验有一定的影响。由实验结果可知,在检测率和误报率之间,如何采取一个折中的方案,使得检测率相对较高,而误报率也相对较低,还有待研究。根据智能手机数据有时序性的特点,还可以运用基于时序的入侵检测技术进行本文实验,可能会有更好的结果。未来在智能手机的入侵检测方面,还需进一步研究探索。

参考文献:

[1] 周忠军,苏红旗,Android 智能手机入侵检测系统设计[J].科技资讯,2012(18):30-32.

[2] Anderson J P.Computer security threat monitoring and surveillance[R].J P Anderson Co,1980.

(下转79页)