基于FWKN-SVM的Android异常入侵检测的研究

孙 敏 徐彩霞 高 阳
(山西大学计算机与信息技术学院 太原 030006)

摘 要 针对Android手机平台提出了基于特征加权K最近邻支持向量机(FWKN-SVM)的异常入侵检测方法。首先,分析了传统SVM在实际应用中的局限性,提出了一种基于特征类内类间距离的特征加权K最近邻的训练集约减策略。随后,根据手机恶意软件对系统造成的影响定义了系统行为,并通过在Android手机上编写的数据采集模块构建测试集和训练集。最后,利用特征加权K最近邻方法进行SVM训练集的精简和分类器的构建,并进行测试集预测。仿真结果表明,FWKN-SVM分类方法在Android异常入侵检测中应用效果良好。

关键词 Android,支持向量机,K最近邻,特征加权,训练集约减,恶意软件
中图法分类号 TP181 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2015.4.022

Research of Android Abnormal Intrusion Detection Based on Feature-weighted K-nearest-neighbor SVM

SUN Min XU Cai-xia GAO Yang
(Department of Computer and Information Technology, Shanxi University, Taiyuan 030006, China)

Abstract In this paper, an abnormal intrusion detection method based on FWKN-SVM (Feature-weighted K-nearest-neighbor Support Vector Machine) for the Android platform was proposed. Firstly, we analyzed the limitations of the traditional SVM in practical applications, and proposed the feature-weighted K-nearest-neighbor method to lessen training set. Then, the system behavior was defined, according to the impact of mobile malware on the system, and the test set and the training set were built by using the data acquisition module implemented on Android phone. Lastly, we used the feature-weighted K-nearest neighbor method to lessen the training set and construct SVM classifier, and then predicted the test set. Simulation result shows that FWKN-SVM classification method has a good performance in Android abnormal intrusion detection.

Keywords Android, SVM, K-nearest-neighbor, Feature-weighted, Lessen training set, Malware

1 引言

支持向量机(Support Vector Machine,SVM)是1995年由Cortes和Vapnik首度提出的一种新型统计学习理论,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,通过在经验风险和置信风险之间寻求最佳折衷,来获得最好的分类和推广能力[1]。SVM有很多其他方法不可比拟的优势:(1)得到的最优解是小样本下的最优解,而不是样本数趋于无穷大时的最优解;(2)算法可以转化为一个凸二次规划问题,理论上可以得到唯一的全局最优解;(3)算法将实际中线性不可分的问题通过核函数巧妙地转换到高维特征空间,以解决原空间中复杂的线性不可分问题[2]。入侵检测是一个典型的模式识别问题,通过检测把正常数据和异常数据区分开。SVM提供了一种解决入侵检测问题的很好的思路,并在理论和实际应用中取得了重要成果。

尽管如此,将SVM应用于智能手机入侵检测的研究还不多。福州大学数学与计算机科学学院的莫宇祥等人采用SVM二分类算法,根据Android OS权限机制及用户自定义策略的角色安全提出一种异常检测模型[3]。它是针对于软件静态权限的一种方法,不能很好地解决软件运行中的动态威胁,而且不能很好地检测出本身具有威胁而权限表现正常的恶意软件。

本文在此基础上,对Android手机进行分析,得出合适的系统运行中的动态描述特征,并提出了一个适用于Android异常入侵检测的特征加权K最近邻支持向量机(FWKN-SVM)。

2 支持向量机(SVM)

针对两类线性可分问题,设样本x是n维向量,即$x_i \in R^n$,y表示这些样本所属类别(正负类)的信息,即$y_i \in \{1,-1\}$。则包含k个样本的训练样本集可以表示为:$(x_i, y_i) \in R^n \times \{+1,-1\}$,其中$i=1,2,\cdots,k$。分类器的构建就转化为在$n$维输入空间上寻找一个能把两类数据正确分开,并使得分类间隔最大的最优分类面的问题。考虑到可能存在一些样本不能被分类面正确分类,为其引入松弛变量ζ_i和惩罚因子C,公式描述如下:

到稿日期:2014-05-08 返修日期:2014-08-08 本文受山西省科技基础条件平台建设项目(2014091004-0105),山西省高等学校教学改革重点项目(J2013010)资助。

孙 敏(1966-),女,副教授,主要研究方向为计算机网络、信息安全,E-mail:minsun@sxu.edu.cn;徐彩霞(1990-),女,硕士,主要研究方向为网络安全、入侵检测、SVM;高 阳(1988-),男,硕士,主要研究方向为移动互联网安全、支付安全。

$$\min \frac{1}{2}\|w\|^2+C\sum_{i=1}^{k}\zeta_i$$

$$\text{s.t. } y_i[w\cdot x_i+b]\geqslant 1-\zeta_i, i=1,2,\cdots,k$$

$$\zeta_i\geqslant 0$$

利用 Lagrange 乘子法，引入拉格朗日乘子 α，将原问题对 ω 和 b 的求解转化为求解 α 的对偶问题，公式描述如下：

$$\max \sum_{i=1}^{k}\alpha_i-\frac{1}{2}\sum_{i,j=1}^{k}\alpha_i\alpha_j y_i y_j(x_i\cdot x_j)$$

$$\text{s.t. } \sum_{i=1}^{k}\alpha_i y_i=0$$

$$\alpha_i\geqslant 0, i=1,2,\cdots,k$$

解上述问题后得到的线性分类函数是：

$$f(x)=\text{sgn}[w\cdot x+b]=\text{sgn}[\sum_{i=1}^{s}\alpha_i y_i(x_i\cdot x)+b]$$

其中，s 为支持向量的个数。对于未知属性的向量 x，可以采用该判决函数来判定其所属类别。

在输入空间非线性可分的情况下，统计学习理论通过使用满足 Mercer 条件的核函数 $K(x_i\cdot x_j)$ 将 n 维输入空间变换到一个高维的特征空间，然后在特征空间中构造最优分类面实现分类，对应的对偶问题为：

$$\max \sum_{i=1}^{n}\alpha_i-\frac{1}{2}\sum_{i,j=1}^{n}\alpha_i\alpha_j y_i y_j K(x_i\cdot x_j)$$

$$\text{s.t. } \sum_{i=1}^{k}\alpha_i y_i=0$$

$$0\leqslant\alpha_i\leqslant C, i=1,2,\cdots,k$$

相应地，最终决策函数为：

$$f(x)=\text{sgn}[\sum_{i=1}^{s}\alpha_i y_i K(x_i\cdot x)+b]$$

3 基于特征加权 K 最近邻的支持向量机

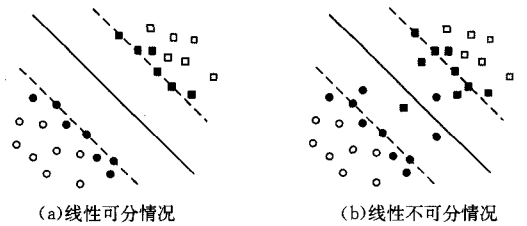
传统 SVM 在解决小样本分类问题时表现出了良好的性能，但是在面对大样本分类问题时表现出了训练时间长、内存开销大的缺点。这是因为 SVM 算法的求解依赖于 Hessian 核矩阵的计算和存储，而核矩阵的规模与样本数量密切相关，当样本数目大的时候，需要很大的内存存储核矩阵，同时核矩阵的计算也需耗费很长的时间[4]。

而在入侵检测的应用中，往往面对的都是成千上万的样本，如果直接使用传统 SVM，就使得问题解决起来很费时或者根本无法解决。

针对于此，本文使用了基于特征加权 K 最近邻的方法来约减 SVM 训练集，大大减少了参与分类器构建的样本个数，从而减少了训练时间和存储规模。

3.1 K 最近邻 SVM

通过分析 SVM 支持向量的分布情况可知，只有在两类样本的相对边界上的边界样本才有可能成为支持向量[5]。如图 1 所示，圆形代表正类样本，方形代表负类样本，实心代表边界向量，在虚线上的实心是支持向量。



(a)线性可分情况　　(b)线性不可分情况

图 1　支持向量和边界向量分布图

如图 1 所示，边界向量集虽然只是训练样本集的一小部分，但是它包含了所有的支持向量。因此，用边界向量集代替全部训练样本进行 SVM 训练，可以在保证 SVM 的分类能力的前提下大幅提高训练速度。

具体地，K 最近邻 SVM 的算法思想如下：选取与某个正类样本 x^+ 距离最近的 K 个负类样本 x^-，认为这部分负类样本属于负类的边界向量。遍历所有的正类样本，找到每个正类样本的 K 个最近邻负类样本，并在剔除重复样本后将其纳入到负类的边界向量集 BV^- 中。同理，遍历所有的负类样本的 K 最近邻样本可以找到正类的边界向量集 BV^+。最后，合并正负类的边界向量集，得到最终的边界向量集，并用该边界向量集作为精简训练集进行 SVM 分类器的构建。

3.2 特征加权 K 最近邻的 SVM

标准 K 最近邻算法认为所有特征对距离影响的程度是等同的，样本间距离是根据样本的所有按相同度量计算的。这种机制存在着弊端，因为有可能由与分类相关性较小的特征或不相关特征对分类结果造成误导，这种现象被称作“维数陷阱”[6]。K 最近邻法对这个问题特别敏感。举例说明：假设每个样本有 20 个特征，在所有这些特征中有 6 个属性与分类强相关、6 个属性与分类弱相关、8 个属性与分类无关。在这种情况下，如果两个样本的 6 个强相关特征的特征值相近，而 6 个弱相关和 8 个无关特征的特征值相差较大，则计算得到的两个样本的距离仍然较大，这就造成分类结果的误差。因此，考虑依据特征与分类的相关性大小对其分配权值，来消弱这种分类误差。强相关特征施加较大的权值，弱相关特征施加较小的权值。

本文使用一种基于类间类内距离的特征重要性评价准则，它能够有效地衡量特征在二分类问题中对于分类贡献的大小，即特征在类别间辨别能力的大小。给定训练样本集 $\{(x_i,y_i)\,|\,x_i\in R^m, y_i\in\{-1,+1\}, i=1,2,\cdots,n\}$，其中 m 表示特征个数，n 表示样本的个数。正类和负类的样本个数分别为 n^+ 和 n^-，则样本的第 i 个特征的区分度计算公式定义如下：

$$F_i=\frac{(\widetilde{x_i^{(+)}}-\widetilde{x_i})^2+(\widetilde{x_i^{(-)}}-\widetilde{x_i})^2}{\frac{1}{n^+-1}\sum_{k=1}^{n^+}(x_{k,i}^{(+)}-\widetilde{x_i^{(+)}})^2+\frac{1}{n^--1}\sum_{k=1}^{n^-}(x_{k,i}^{(-)}-\widetilde{x_i^{(-)}})^2}$$

其中，$\widetilde{x_i}$、$\widetilde{x_i^{(+)}}$、$\widetilde{x_i^{(-)}}$ 分别表示第 i 个特征在整个样本集、正类样本集和负类样本集上的均值，$x_{k,i}^{(+)}$ 表示正类第 k 个样本点的第 i 个特征的特征值，$x_{k,i}^{(-)}$ 表示负类第 k 个样本点的第 i 个特征的特征值。F_i 值越大，表示第 i 个样本对分类的贡献越大，辨识能力越强。计算出每个特征的特征权重后，在计算样本之间欧氏距离时使用加权的欧氏距离公式：

$$d(x_i,x_j)=$$

$$\sqrt{w_1(x_{i1}-x_{j1})^2+w_2(x_{i2}-x_{j2})^2+\cdots+w_m(x_{im}-x_{jm})^2}$$

这里存在一个问题，因为 SVM 中引入了核函数解决线性不可分问题，计算两个样本之间的距离时应该是计算通过核函数映射到高维特征空间中的样本距离，即核距离，而我们提出的加权欧氏距离映射到高维空间中并不好求解。

但是，根据 Burges 提出的“对某些特定的 kernel(核函数)，邻域内点之间的相对距离在输入空间和特征空间中保持不变”[7]的性质，我们可以用在输入空间中计算加权欧氏距离求出的 K 最近邻近似代替特征空间中的样本 K 最近邻。对于 RBF 核，上述性质很显然是成立的，证明如下。

已知两个样本 x_i，x_j，分别计算它们在输入空间和特征

空间的距离 $d(x_i,x_j)$和 $d'(x_i,x_j)$：

$$d(x_i,x_j)=\sqrt{\|x_i-x_j\|^2}$$

$$d'(x_i,x_j)=\sqrt{\|\varphi(x_i)-\varphi(x_j)\|^2}$$

$$=\sqrt{k(x_i,x_i)+k(x_j,x_j)-2k(x_i,x_j)}$$

$$=\sqrt{\exp(-r\|x_i-x_i\|^2)+\exp(-r\|x_j-x_j\|^2)-2\exp(-r\|x_i-x_j\|^2)}$$

$$=\sqrt{2\exp(0)-2\exp(-r\|x_i-x_j\|^2)}$$

$$=\sqrt{2}\cdot\sqrt{1-\frac{1}{e^{r\|x_i-x_j\|^2}}}$$

由于 $e^{r\|x_i-x_j\|^2}$（$r\in R^+$）是单调递增的，因此 $d(x_i,x_j)$增大，$d'(x_i,x_j)$随之增大；$d(x_i,x_j)$减小，$d'(x_i,x_j)$随之减小。因此，在核空间和输入空间中求出的 K 最近邻应该相同。下面仿真实验的结果，也能说明该理论的正确性。

因此，本文的支持向量机选择 RBF 核函数，并在输入空间中根据加权 K 最近邻思想求解边界向量集。

基于加权 K 最近邻的 SVM 的主要步骤如下：

1）计算每个特征的特征权重 w_i；

2）在输入空间中利用基于加权欧氏距离的 K 最近邻思想求解训练样本集的正负边界向量集，并合并组成总的边界向量集；

3）利用基于 RBF 核函数的 SVM，训练步骤 2）得到的边界向量集，得出分类器。

4）对测试样本用步骤 3）得到的分类器进行预测。

4 Android 入侵检测数据集构建

系统行为是指能够表示系统状态的一组信息，用于区分系统的正常和异常状态。系统行为的选择对异常检测而言至关重要[8]。通过分析现有的大多数恶意软件对系统造成的影响，选择受影响较大的信息作为定义系统行为所需的信息。目前，手机恶意软件的恶意行为主要分为以下 4 方面：恶意扣费、隐私窃取、垃圾短信和系统破坏[9]。这 4 方面的恶意行为对系统造成的影响如表 1 所列。

表 1　恶意行为造成的影响

恶意行为	行为描述	受影响较大的系统信息
恶意扣费	私自发送短信定制费用高昂的 SP 服务，并自动屏蔽 10086 开头的全部短信，使用户无法察觉，或者是私自下载文件，消耗用户流量	短信、流量、内存、SD 卡占用情况
隐私窃取	窃取用户的通讯录、照片、短信、财务等信息，上传至服务器	短信、流量
垃圾短信	包括打折促销、地产、理财等服务类短信和冒充亲友诈骗、中奖钓鱼诈骗等恶意欺诈类短信	短信
系统破坏	开机自启动、后台运行、自行提高应用程序权限，造成内存、CUP 等的异常，甚至可能造成系统崩溃	内存、CUP、进程

综上分析，本文选取了 17 个关键特征属性作为系统行为，包括 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。上述均为连续型数值属性。在 Android 平台上编程实现数据采集模块来提取在正常和受攻击状态下的样本，分别为正常样本和异常样本，在正、异常样本中分别随机抽取 80%的样本组成训练集，剩下的样本构成测试集。

5 仿真实验

通过实验研究加权 K 最近邻 SVM 本身的性能，以及它在 Android 入侵检测中的适用情况。实验平台为 Intel 2.94 GHz CPU，2GB RAM，Windows 7 操作系统，所有算法均采用 java 语言在 Eclipse 平台上实现。实验时设定参数 $K=3$，SVM 的核函数采用径向基（RBF）核函数：$K(x_i,x_j)=\varphi(x_i)\cdot\varphi(x_j)=\exp(-r\|x_i-x_j\|^2)$，并采用交叉验证的方法来选择训练参数，结果为 $C=10$，$r=0.15$。

实验提出的算法在 3 个广泛应用于分类领域的 UCI 标准数据集和一个 Android 入侵检测数据集上进行检验（为了使数据特征之间具有可比性，已对数据集进行了标准化处理）。

5.1 UCI 数据集仿真实验

选用 Wine、Abalone 和 Poker-hand 3 个规模不同的数据集来进行实验，将每个数据集分成两类样本数据，如表 2 所列。

表 2　UCI 数据集

数据集	训练集样本数		测试集样本数	属性数
	正类	负类		
Wine	115	36	27	13
Abalone	1341	1159	335	8
Poker-hand	2421	2579	20010	10

在本实验中同时采用了标准 SVM 算法（LIBSVM）、K 最近邻 SVM（KNN-SVM）和基于核距离的 KNN-SVM（KKNN-SVM）对数据集进行训练分类，并与本文中的加权 K 最近邻 SVM（WKNN-SVM）进行了比较。为了使结果具有可比性，4 种算法中的参数选择保持不变。具体的实验结果如表 3 所列，其中训练样本个数是指实际参加 SVM 分类器构建的样本个数，训练时间是包括了边界向量集生成时间的总时间。

表 3　UCI 数据集实验结果比较

数据集	分类算法	训练样本/个	约减率/%	支持向量/个	准确率/%	训练时间/s
Wine	LIBSVM	151		15	86.88	0.020
	KNN-SVM	31	79.5	12	85.18	0.003
	KKNN-SVM	31	79.5	12	85.18	0.003
	WKNN-SVM	16	89.4	10	86.88	0.002
Abalone	LIBSVM	2500		2295	49.85	3.054
	KNN-SVM	2129	14.84	2063	49.25	0.716
	KKNN-SVM	2129	14.84	2063	49.25	0.716
	WKNN-SVM	2141	14.36	2044	49.55	0.774
Poker-hand	LIBSVM	5000		4569	49.9	15.43
	KNN-SVM	3607	28.86	3337	49.98	5.591
	KKNN-SVM	3607	28.86	3337	49.98	5.591
	WKNN-SVM	3621	28.58	3294	50.01	5.762

注：C、r 和 K 参数取其他值时，实验结果仍然呈上述规律，即同样参数条件下 K 最近邻 SVM 的训练时间比 LIBSVM 短，而准确率基本保持不变。

分析上述实验结果，可以得出以下几个结论：

第一，使用 RBF 核函数求解出的边界向量集与普通欧氏距离求出的边界向量集完全相同，这可从 KNN-SVM 和 KKNN-SVM 的对比结果看出，从而再次验证了文章中提到的“对 RBF 核函数，邻域内点之间的相对距离在输入空间和特征空间中保持不变”的说法。

第二，使用 K 最近邻思想对支持向量机减样的方法可以

（下转第 131 页）

DDoS检测机制及性能分析[J]. 软件学报,2012,23(5):1272-1280
[4] 张永铮,肖军,云晓春,等. DDoS攻击检测和控制[J]. 软件学报,2012,23(8):2258-2072
[5] 王睿. 一种基于回溯的Web上应用层DDOS检测防范机制[J]. 计算机科学,2013,40(11A):175-177
[6] 夏秦,王志文,卢柯. 入侵检测系统利用信息熵检测网络攻击的方法[J]. 西安交通大学学报,2013,47(2):14-19
[7] 周华,周海军,马建锋. 基于博弈论的入侵容忍系统安全性分析模型[J]. 电子与信息学报,2013,35(8):1933-1939
[8] Bimal K M,Gholam M A. Differential epidemic model of virus and worms in computer network [J]. International Journal of Network Security,2012,14(3):149-155
[9] Zhu Q Y, Yang X F, Yang L X, et al. Optimal control of computer virus under a delayed model [J]. Applied Mathematics and Computation,2012,218(23):11613-11619
[10] 张辉. 自体集网络入侵检测中的高效寻优算法仿真[J]. 计算机仿真,2013,30(8):297-300
[11] 樊爱宛,时合生. 基于特征选择和SVM参数同步优化的网络入侵检测[J]. 北京交通大学学报,2013,37(5):58-61
[12] 饶雨泰,杨凡. 网络入侵搅动下的网络失稳控制方法研究[J]. 科技通报,2014,30(1):185-188
[13] 罗柏文,沈彩耀,于宏毅. 采用余弦调制滤波器组的多径衰落信号子带合成[J]. 信号处理,2013,29(5):537-543

(上接第118页)

有效地约减训练集,可以在保证分类精度基本不变的前提下大大减少训练时间;在Poker-hand数据集上约减后的SVM的分类精度反而提高了,这是由于减样的过程把部分噪声数据也除掉了。

第三,加权K最近邻方法对训练集的约减效果比普通K最近邻方法效果更好。对比WKNN-SVM和KNN-SVM可知,WKNN-SVM比KNN-SVM的分类精度更好,这是一种更有效的减样方法。

5.2 Android数据集仿真实验

使用上文提到的Android数据集构建方法,采集了不同时间段中3个规模不同的数据集,如表4所列。

表4 Android数据集描述

数据集	训练集样本数		测试集样本数
	正类	负类	
Android-1	587	93	170
Android-2	2600	684	821
Android-3	4886	924	1450

本实验将本文提出的FWKN-SVM方法和传统的SVM方法(LIBSVM)进行对比,实验选取相同的参数。准确率的对比结果如图2所示,训练时间的对比结果如图3所示。

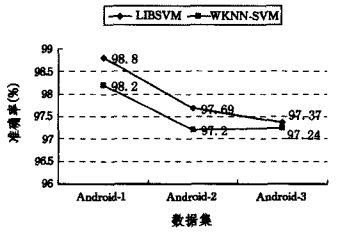


图2 准确率对比图

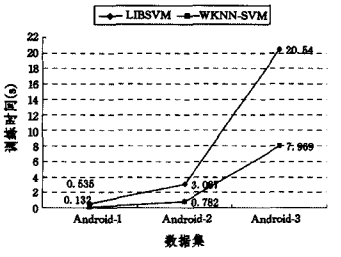


图3 训练时间对比图

上述结果表明:

第一,本文采用的数据集采集和构建方法是有效的,是适用于Android手机SVM入侵检测研究的。因为使用两种方法得到的准确率都比较高,在实时应用中的效果比较好。

第二,WKNN-SVM比SVM更适合于Android入侵检测,在保证准确率和误报率基本不变的前提下大大减少了训练时间,而且,训练样本集越大,训练时间减少得越明显,约减的效果越明显。

结束语 本文提出了一种基于加权K最近邻支持向量机的Android手机入侵检测方法。该方法通过分析恶意软件对系统造成的影响定义了Android入侵检测系统行为,并进行数据集构建;考虑到各个特征值对分类结果的不同影响,使用了基于类内类间距离的特征加权方法求解样本的K最近邻,进而得出训练样本边界向量集;针对得到的边界向量集合进行支持向量机训练,可以在保证分类精度不变的前提下提高训练速度并减少内存占用,适用于实际应用中的大规模样本分类问题。因此,本文的方法为解决Android手机异常入侵检测提供了一种思路。

参 考 文 献

[1] Vapnik V. The nature of statistical learning theory [M]. Springer,2000
[2] 钱权,耿焕同,王煦法. 基于SVM的入侵检测系统[J]. 计算机工程,2006,32(9):136-138
[3] 莫宇祥,俞建銮,王磊,等. 基于角色的Android手机平台木马检测系统[J]. 现代计算机:上半月版,2012(12):51-55
[4] 罗瑜,易文德,王丹琛,等. 大规模数据集下支持向量机训练样本的缩减策略[J]. 计算机科学,2007,34(10):211-213
[5] 孙发圣,肖怀铁. 基于K最近邻的支持向量机快速训练算法[J]. 电光与控制,2008,15(6):44-47
[6] 陈振洲,李磊,姚正安. 基于SVM的特征加权KNN算法[J]. 中山大学学报:自然科学版,2005,44(1):17-20
[7] Burges C J C. Geometry and invariance in kernel based methods [M]// Advances in Kernel Methods-Support Vector Learning. Cambridge,MA:MIT Press,1998:89-116
[8] 乜聚虎. 智能手机异常检测技术研究与实现[D]. 合肥:中国科学技术大学,2011
[9] 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