基于支持向量机的入侵检测系统*

饶 鲜+, 董春曦, 杨绍全

(西安电子科技大学 电子工程系 电子对抗研究所,陕西 西安 710071)

An Intrusion Detection System Based on Support Vector Machine

RAO Xian+, DONG Chun-Xi, YANG Shao-Quan

(Institute of Electronic Counter Measures, Department of Electronics Engineering, Xidian University, Xi'an 710071, China)

+ Corresponding author: Phn: 86-29-8202274, E-mail: xianrao@yahoo.com.cn

http://ecm.xidian.edu.cn

Received 2001-12-10; Accepted 2002-08-02

Rao X, Dong CX, Yang SQ. An intrusion detection system based on support vector machine. *Journal of Software*, 2003,14(4):798~803.

Abstract: The generalizing ability of current IDS (intrusion detection system) is poor when given less priori knowledge. Utilizing SVM (support vector machines) in Intrusion Detection, the generalizing ability of IDS is still good when the sample size is small (less priori knowledge). First, the research progress of intrusion detection is recalled and algorithm of support vector machine taxonomy is introduced. Then the model of an Intrusion Detection System based on support vector machine is presented. An example using system call trace data, which is usually used in intrusion detection, is given to illustrate the performance of this model. Finally, comparison of detection ability between the above detection method and others is given. It is found that the IDS based on SVM needs less priori knowledge than other methods and can shorten the training time under the same detection performance condition.

Key words: intrusion detection; network security; support vector machine; statistical learning; pattern recognition

摘 要: 目前的入侵检测系统存在着在先验知识较少的情况下推广能力差的问题.在入侵检测系统中应用支持向量机算法,使得入侵检测系统在小样本(先验知识少)的条件下仍然具有良好的推广能力.首先介绍入侵检测研究的发展概况和支持向量机的分类算法,接着提出了基于支持向量机的入侵检测模型,然后以系统调用执行迹(system call trace)这类常用的入侵检测数据为例,详细讨论了该模型的工作过程,最后将计算机仿真结果与其他检测方法进行了比较.通过实验和比较发现,基于支持向量机的入侵检测系统不但所需要的先验知识远远小于其他方法,而且当检测性能相同时,该系统的训练时间将会缩短.

关键词: 入侵检测;网络安全;支持向量机;统计学习;模式识别

中图法分类号: TP181 **文献标识码**: A

* Supported by the Military Communication Pre-Research Project of the 'Tenth Five-Year-Plan' of China under Grant No. 4100104030 ("十五"军事通讯预研)

第一作者简介: 饶鲜(1976－),女,陕西城固人,博士生,讲师,主要研究领域为网络安全,信息对抗.

随着计算机和网络技术应用的日益普及,计算机网络安全越来越受到人们的重视.入侵检测作为网络安全研究的重要内容,更是引起了国内外学者的广泛关注.入侵检测(intrusion detection)通过检查有关的审计数据,例如系统日志或网络数据,以判断系统中是否有违背安全策略或计算机系统安全的行为.

入侵检测可以看作是一个分类问题,也就是对给定的审计数据进行分类:什么样的数据是正常的,什么样的数据是异常的.在有关文献中,Forrest[1]等人把入侵检测看作是区分“自我”(也就是“正常”)和“非自我”(也就是“异常”)的过程,提出了基于免疫模型的入侵检测技术.Ghosh[2]利用神经网络来提取特征和分类.W.Lee[3]从数据挖掘技术的角度探讨了入侵检测的实现问题.以上方法都需要大量或者是完备的审计数据集才能达到比较理想的检测性能,并且训练时间较长.那么,如何在小样本的情况下,提取审计数据特征,实现入侵检测呢?

支持向量机(support vector machines)是一种建立在统计学习理论基础之上的机器学习方法.其最大的特点是根据 Vapnik[4]结构风险最小化原则,尽量提高学习机的泛化能力,即由有限的训练集样本得到小的误差仍然能够保证对独立的测试集保持小的误差.另外,由于支持向量算法是一个凸优化问题,所以局部最优解一定是全局最优解.这是其他学习算法所不及的.

将支持向量机应用到入侵检测中,可以保证在先验知识不足的情况下,支持向量机分类器仍有较好的分类正确率,从而使得整个入侵检测系统(intrusion detection system)具有较好的检测性能.由此,我们提出了一种基于支持向量机的入侵检测模型.

本文首先介绍了支持向量机的基本原理,然后提出了基于支持向量机的入侵检测系统的模型,最后以系统调用执行迹作为入侵检测数据,详细讨论了该模型的工作过程,并给出了计算机仿真的结果.

1 支持向量机的分类算法

对于分类问题,支持向量机算法根据区域中的样本计算该区域的决策曲面,由该曲面确定该区域中样本的类别.由于入侵检测可以看成是一个二分类的问题,这里我们主要讨论计算决策面和对样本分类的方法.

1.1 线性支持向量机

样本 x 为 k 维向量,在某区域内的 l 个样本所属类别为 $(x_1,y_1),...,(x_l,y_l)\in R_k\times\{\pm1\}$.若超平面:

$$w\cdot x+b=0 \tag{1}$$

能将样本分为两类,其中·表示向量的点积.最佳的超平面应使两类样本到超平面的距离为最大.显然,式(1)中的 w 和 b 乘以系数以后仍能满足方程.不失一般性,对于所有的样本 x_i ,式 $|w\cdot x_i+b|$ 的最小值为 1,则样本与此最佳超平面的最小距离为 $|(w\cdot x_i+b)|/\|w\|=1/\|w\|$.最佳超平面应满足约束:

$$y_i[(w\cdot x_i)+b]\geq 1,\ i=1,...,l. \tag{2}$$

w 和 b 的优化条件应该是使两类样本到超平面最小距离之和 $2/\|w\|$ 最大.另外,考虑到可能存在一些样本不能被超平面正确分类,因此引入松弛变量:

$$\xi_i\geq 0,\ i=1,...,l, \tag{3}$$

问题变成在式(2)和式(3)的条件下最小化

$$\frac{1}{2}\|w\|^2+C\cdot\sum_{i=1}^{l}\xi_i, \tag{4}$$

其中 C 为一个正常数.上式的第 1 项使样本到超平面的距离尽量大,从而提高泛化能力;第 2 项则使误差尽量小.利用 lagrange 乘子法,可以把式(4)变成其对偶形式,从而有

$$\begin{aligned}&\max W(a)=\sum_{i=1}^{l}\alpha-\frac{1}{2}\sum\alpha_i y_i\alpha_j y_j(x_i\cdot x_j),\\ &\text{s.t.}\quad \sum_{i=1}^{l}\alpha_i y_i=0,\\ &\qquad \alpha_i\in[0,C],\ i=1,...,l,\end{aligned} \tag{5}$$

以及

$$w=\sum_{i=1}^{l}\alpha_i y_i x_i. \tag{6}$$

这是一个典型的二次优化问题,已由高效的算法求解.可以证明,在此优化问题的解中有一部分 α_i 不为 0,它们所对应的训练样本完全确定了这个超平面,因此称其为支持向量.按照优化理论的 Kuhn-Tucker 定理,在鞍点,对偶变量与约束的乘积为 0,从而求得超平面的另一个参数 b 满足:

$$y_i(w\cdot x_i+b)=1. \tag{7}$$

对于未知属类的向量 x, 可以采用线性判决函数

$$f(x)=\mathrm{sgn}(w\cdot x+b) \tag{8}$$

来判定其所属类别.综合式(6),得到:

$$f(x)=\mathrm{sgn}\left(\sum_{i=1}^{l}\alpha_i y_i(x_i\cdot x)+b\right). \tag{9}$$

1.2 非线性支持向量机

在线性支持向量机的训练算法中,数据以点积的形式 $(x_i\cdot x_j)$ 出现.现在用非线性映射把输入空间映射到某一特征空间,记为 $\Psi: R_k\to H$.如果存在一种核函数 K,使得

$$K(x_i,x_j)=\left(\Psi(x_i)\cdot\Psi(x_j)\right) \tag{10}$$

可以在特征空间中进行许多计算,而不需要知道具体的映射 Ψ .

现在可用核函数代替线性支持向量机中的点积形式,式(5)的对偶规划可变为

$$\begin{aligned}&\max W(a)=\sum_{i=1}^{l}\alpha-\frac{1}{2}\sum\alpha_i y_i\alpha_j y_j K(x_i\cdot x_j),\\ &\text{s.t.}\quad \sum_{i=1}^{l}\alpha_i y_i=0,\\ &\qquad \alpha_i\in[0,C],\ i=1,\ldots,l.\end{aligned} \tag{11}$$

非线性支持向量机的判决函数为

$$f(x)=\mathrm{sgn}\left(\sum\alpha_i y_i K(x_i,x)+b\right). \tag{12}$$

2 基于支持向量机的入侵检测系统

基于支持向量机的入侵检测系统主要由审计数据预处理器、支持向量机分类器和决策系统 3 部分组成,如图 1 所示.

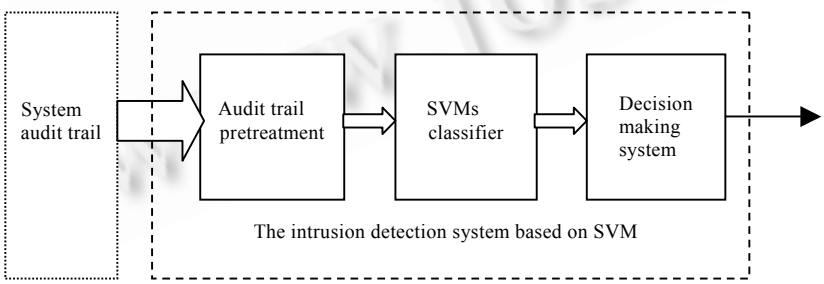

Fig.1 The intrusion detection system based on SVM
图 1 基于支持向量机的入侵检测系统

审计数据预处理器用来对大量的系统审计纪录进行处理或变换.由于支持向量机的分类器只能对维数相同的数字向量进行分类,但系统审计数据中的数据不但长度不尽相同,而且很有可能不是数字类型,所以必须将原始数据转换成支持向量机能够识别的数字向量.支持向量机分类器对这些数字向量进行分类,产生判决结果.当然,这些判决结果可以直接作为整个入侵检测系统的输出,但为了进一步提高整个系统的正确率,我们可以设定一些判决准则,例如发生数目、百分比等来进行最终的判定.这个过程是由决策系统完成的.

整个系统的工作过程分为两个阶段:训练阶段和检测阶段.在训练阶段,根据已知的正常审计数据和异常审计数据按照式(11)来训练支持向量机,并根据式(6)和式(7)得到支持向量和相应的参数.在检测阶段,预处理器先将未知状态的审计数据处理成数字向量的形式,然后通过支持向量机分类器,根据判决函数式(12)对这些数字向量进行分类,并将分类结果提交给决策系统作出最后的判断.

3 实验仿真

可用于入侵检测的数据类型很多,这里我们选用系统调用序列进行仿真.Forrest 等人在研究中发现:系统关键程序的执行,可以通过程序执行过程中所使用的系统调用序列(也称为执行迹(trace))来描述.一个正常行为可以由其执行迹局部模式,即系统调用短序列来描述,其程序执行代码具有相对的稳定性.在异常行为中,可能出现和正常情况有一定的差别的系统调用短序列.也就是说,在正常的系统调用中出现的都是正常的短序列,而异常的执行迹中除了正常的短序列以外,还会出现异常的短序列.判断该执行迹是正常的还是异常的就转化为识别执行迹中短序列是正常的还是异常的.这里,我们使用支持向量机来实现对短序列的分类.

在下面的计算机仿真中,我们选用 MIT(Massachusetts Institute of Technology)人工智能实验室(AI Lab.)公开提供的 lpr 数据进行仿真,以此为例来详细讨论基于支持向量机的入侵检测系统的工作过程.

3.1 数据预处理和短序列长度的选择

该数据集中的每个执行迹数据文件由两列数据构成,第 1 列为进程标识符,第 2 列为进程的系统调用命令在系统调用名称列表(mapping file)中的索引值,如‘5’代表‘open’.进程标识符相同的系统调用构成一个进程的执行迹.因为该数据已是数字序列,预处理的主要目的是得到该执行迹的系统调用短序列.具体做法是用长度为 k 的窗口在程序执行迹上滑动来得到这个执行迹的短序列.那么,长度 k 选取何值是最合适的?

系统调用短序列反映了进程执行过程中系统调用之间的次序关系.如果选取的短序列长度为 1,就丢掉了系统调用的次序信息,而长度太大,就丢掉了系统执行的局部信息状况,无法正确反映正常和异常情况下的局部序列调用状况.W.Lee[5]教授从信息论的角度研究了数据长度的选择情况,认为最合适的系统调用短序列长度为 6~7.本实验选取系统调用短序列的长度为 6.

3.2 获得训练样本

支持向量机的参数是通过训练得到的,所以需要获得两类训练样本,即正常短序列样本和异常短序列样本.

我们用长度为 k 的滑动窗口对已知正常的系统调用执行迹进行扫描,可以得到正常的系统调用短序列样本.由于入侵的非法活动只占程序执行的一小部分,所以异常短序列只占异常执行迹的很小一部分.当我们用长度为 k 的滑动窗口对于异常的执行迹进行扫描时,会得到一组既有正常短序列又有异常短序列的系统调用短序列列表.将这组短序列列表与已获得的正常短序列样本进行比较,不同于正常短序列的那些系统调用短序列就构成了异常短序样本.如果设我们得到了 l 个样本,可将所有的短序列样本记为 $(x_1, y_1),...,(x_l, y_l) \in R_k \times \{\pm 1\}$ 的形式,其中正常样本的标签 y_i 为+1;异常样本 y_i 为−1,且 $k = 6$.所得到的正常和异常的系统调用短序列训练数据样本见表 1.

Table 1 Example of the training samples of normal and abnormal short sequences

表 1 正常和异常的系统调用短序列训练样本实例

Short sequences (with length 6) ($x_i \in R_6$)	Classify label(y_i)
5 3 67 67 5 139	“normal”(+1)
3 67 67 5 139 67	“normal”(+1)
…	…
19 4 6 9 10 6	“abnormal”(−1)
4 6 9 10 6 6	“abnormal”(−1)
…	…

3.3 决策准则

由于我们获得训练样本时所得到的正常短序列样本并不完备,导致在基于正常短序列上获得的异常短序

列样本中可能包含正常的短序列,从而使得支持向量机的分类器产生一些分类错误,并且支持向量机分类器本身也具有不精确性,所以我们需要设定某些判断规则来提高整个检测系统的性能.这些规则主要有两种.一种是根据异常系统调用短序列的数目进行判断.首先我们选定一个阈值,然后对于给定系统进程执行迹的短序列进行分类,如果异常系统调用短序列的数目超过阈值,则判定为异常;反之,则判定为正常.另一种规则是根据异常系统调用短序列在整个系统调用短序列中所占的百分比进行判断.该方法与第 1 种方法相似,也需要选定阈值,先对给定进程执行迹的短序列进行分类,然后统计异常系统调用短序列所占的百分比.如果百分比大于阈值,则判定为异常;反之则判定为正常.由于以上两种方法简单、有效,所以在系统中常常得到应用.

3.4 入侵检测仿真

正如前面所提到的,基于支持向量机的检测系统进行入侵检测的过程分为两个阶段:训练阶段和检测阶段.

在训练阶段,我们用已获得的训练样本根据式(11)训练支持向量机得到支持向量和相关参数.在检测阶段,我们就可以对未知状态的系统进程执行迹进行判定了.首先,审计数据预处理器先用长度为 k=6 的滑动窗口对该执行迹进行扫描,得到了一组系统调用短序列.将这些系统调用的短序列输入支持向量机分类器,利用式(12)可以得到判决向量 Y.然后根据上面所提到的第 1 种判决规则对该判决向量进行判决,即可判定该执行迹的状态.表 2 给出了本文方法训练数据和检测数据的分配情况.

Table 2 Distribution of the data used in training and the test in the simulation

表 2 仿真中训练数据和检测数据的分配情况

Process	Training data set		Testing data set	
	Number of normal traces	Number of abnormal traces	Number of normal traces	Number of abnormal traces
MIT lpr	20	10	2 704	1 001

表 3 给出了本文的仿真结果与 Christina Warrender[6]等人得出的研究结果的比较.Christina Warrender 等人主要研究了 4 种算法.Stide(sequence time-delay embedding)方法是根据训练数据集预先列出长度为 k 的所有独特、连续的系统调用序列作为正常行为的特征.然后统计所要检测的执行迹的这些特征,并将它们与已建立的正常行为的特征比较,如果不同特征的数目超过阈值就判定有异常发生.t-stide(stide with frequency threshold)与 stide 类似,只是在检测时利用的是不同特征出现的频率作为判决规则.RIPPER(repeated incremental pruning to produce error reduction)是 William Cohen 提出的一种规则学习算法.这种算法主要应用于分类问题.HMM(hidden Markov model)是用隐马尔可夫模型(HMM)对正常的执行迹进行建模.在检测阶段,对于状态转移概率和输出概率给出阈值,如果低于阈值,则判定有异常发生.他们提供的数据是在检测率为 100%的情况下最低虚警率的数据.由表 3 可以看出,HMM 的检测性能最好,但花费的时间较长.对于相同的检测结果,本文的 SVM 所需的训练时间要远远小于 HMM 方法.

Table 3 Comparison of the performances of several intrusion detection methods

表 3 几种入侵检测方法的性能比较

	Stide	t-Stide	RIPPER	HMM	SVM
Time used in training	10 min	10 min	1 min	5 days	7 min
The lowest false alarm	0.0	0.007 5	0.001 6	0.000 3	0.000 3

通过实验可以看出,基于支持向量机的入侵检测模型具有以下特点:首先,它不需要全部的正常和异常的信息,在给出较少的正常和异常执行迹的情况下就能得到比较理想的检测效果;其次,该方法所需要的训练时间和检测时间比其他方法短,所以该方法能够随时升级,并进行高效的实时检测.

4 结 论

支持向量机作为一种对于小样本数具有良好学习性能的机器学习方法,已经被用于网页识别、人脸识别等许多方面.本文初步探讨了它在网络入侵检测方面的应用,提出了基于支持向量机的入侵检测框架和工作原理,并利用系统调用序列(执行迹)进行了仿真实验,所得到的性能令人满意.由于支持向量机方法建立了一套有限样本下机器学习的理论框架和通用方法,能够较好地解决小样本、非线性等实际分类问题,它在网络安全中的应用将会越来越广泛.

致谢 西安电子科技大学电子对抗研究所的杨绍全教授对本文的工作给予了细心的指导,陈卫东博士对本文的完成提出了很多有益的建议,在此一并表示感谢.

References:

[1] Forrest S, Perrelason AS, Allen L, Cherukur R. Self_Nonself discrimination in a computer. In: Rushby J, Meadows C, eds. Proceedings of the 1994 IEEE Symposium on Research in Security and Privacy. Oakland, CA: IEEE Computer Society Press, 1994. 202~212.

[2] Ghosh AK, Michael C, Schatz M. A real-time intrusion detection system based on learning program behavior. In: Debar H, Wu SF, eds. Recent Advances in Intrusion Detection (RAID 2000). Toulouse: Spinger-Verlag, 2000. 93~109.

[3] Lee W, Stolfo SJ. A data mining framework for building intrusion detection model. In: Gong L, Reiter MK, eds. Proceedings of the 1999 IEEE Symposium on Security and Privacy. Oakland, CA: IEEE Computer Society Press, 1999. 120~132.

[4] Vapnik VN. The Nature of Statistical Learning Theory. New York: Spring-Verlag, 1995.

[5] Lee W, Dong X. Information-Theoretic measures for anomaly detection. In: Needham R, Abadi M, eds. Proceedings of the 2001 IEEE Symposium on Security and Privacy. Oakland, CA: IEEE Computer Society Press, 2001. 130~143.

[6] Warrender C, Forresr S, Pearlmutter B. Detecting intrusions using system calls: Alternative data models. In: Gong L, Reiter MK, eds. Proceedings of the 1999 IEEE Symposium on Security and Privacy. Oakland, CA: IEEE Computer Society Press, 1999. 133~145.

第 20 届全国数据库学术会议(NDBC2003)

征 文 通 知

由中国计算机学会数据库专业委员会(www.ccf-dbs.org.cn)主办、中国人民解放军国防科学技术大学(www.nudt.edu.cn)承办的第 20 届全国数据库学术会议将于 2003 年 10 月 10 日~12 日在湖南召开。会议网址为 http://ndbc2003.nudt.edu.cn/。

一、征文范围(但不限于这些领域)

A. Web 与数据库
B. 移动计算和数据库
C. 数据库实现技术
D. 数据库安全性
E. 数据仓库和 OLAP
F. 数据挖掘和知识发现
G. 文本数据库与信息检索
H. 查询处理与用户界面
I. 生物与基因信息系统
J. 数字图书馆
K. 数据集成和迁移
L. 面向对象与对象-关系数据库系统
M. 中间件和应用服务器技术
N. 实时数据库系统
O. 科学与统计数据库
P. XML 和半结构化数据库系统
Q. 空间和时态数据库系统
R. 内容管理/知识管理
S. 工作流/数据库应用
T. 并行和分布式数据库系统
U. 多媒体数据库技术
V. 电子商务/电子政务

二、重要日期

论文提交截止时间：2003 年 5 月 10 日
论文录用通知时间：2003 年 7 月 20 日
排版稿件截止时间：2003 年 8 月 10 日

三、会议咨询

会议信息可以通过访问网站 http://ndbc2003.nudt.edu.cn/得到，也可以与会务组联系。
E-mail: ndbc2003@nudt.edu.cn
电话：86-731-4552710（NDBC2003 会务组：梁妍女士、隋品波女士）
传真：86-731-4512504
通讯地址：湖南省长沙市国防科学技术大学计算机学院 NDBC2003 会务组（邮编 410073）