第34卷 第18期 计 算 机 工 程 2008年9月
Vol.34 No.18 Computer Engineering September 2008
·安全技术· 文章编号：1000—3428(2008)18—0199—03 文献标识码：A 中图分类号：TP393

基于AdaBoost特征约减的入侵检测分类方法

陶晓玲[1]，王 勇[1]，罗 鹏[2]
(1. 桂林电子科技大学网络中心，桂林 541004；2. 广西移动通信有限责任公司，桂林 541004)

摘 要：提出一种基于AdaBoost的入侵特征约减算法，利用该算法约减入侵特征中的冗余特征，构造Ada-加权和Ada-域值分类器，并与支持向量机分类器进行对比。设计并实现Linux实时入侵检测实验平台，并将特征约减算法和3种分类方法应用于该平台。实验结果表明，由特征约减算法挑选出来的入侵特征集较优，Ada-加权和 Ada-域值分类器的分类效果优于支持向量机分类器，且 Ada-域值分类器在测试集上的检测性能最佳。

关键词：入侵检测；特征约减；Ada加权分类器；Ada域值分类器

Classification Method of Intrusion Detection Based on AdaBoost Feature Reduction

TAO Xiao-ling[1], WANG Yong[1], LUO Peng[2]
(1. Network Information Center, Guilin University of Electronic Technology, Guilin 541004;
2. Guangxi Mobile Communication Co., Ltd., Guilin 541004)

【Abstract】 A reduction algorithm based on AdaBoost is proposed in the paper to reduce the intrusion feature redundancy. With algorithm, two classifiers——Ada weighted-classifier and Ada threshold-classifier are constructed, compared with support vector machine classifier. An Linux IDS experimental platform is designed and implemented, and the algorithm and three classification methods are applied using the platform. Experimental results show that intrusion feature set selected by the feature reduction algorithm is better, and the classification effect of Ada weighted-classifier and Ada threshold-classifier are better than SVM classifier, also the performance of detection Ada threshold-classifier is the best on test set.

【Key words】 intrusion detection; feature reduction; Ada weighted-classifier; Ada threshold-classifier

作为网络安全的第二道防线，入侵检测技术是目前研究热点。为了获得理想的入侵检测效果，一方面需要根据主机或网络的原始信息，通过传感器来提取关键入侵特征；另一方面需要建造一个好的分类器，及时发现和识别系统中的入侵行为和企图，给出入侵警报。对于前者，如果能将入侵检测的关键特征找出来，则冗余的特征就可以不必检测，从而降低数据采集的成本、提高采集速度，即特征选择问题。本文将机器学习算法AdaBoost引入异常检测的特征选择和约减中，使选用的输入特征能为分类器提供最有用的信息。对此，人们根据不同的原理构造分类器[1]，试图取得最佳的检测效果。本文在基于AdaBoost的特征约减算法基础上构造了Ada-加权和Ada-域值分类器，取得了比支持向量机更好的分类效果。目前，多数入侵检测方法都是使用KDD 99 数据集进行评估的。然而，面对日益复杂的网络环境，该数据集具有局限性。因此，本文在借鉴麻省理工学院林肯实验室成功经验的基础上，设计并实现了在新的软硬件环境下的Linux实时入侵检测实验环境，用于验证特征约减算法以及Ada-加权、Ada-域值分类算法的有效性。

1 基于 AdaBoost 的入侵特征约减

鉴别并选取重要的输入特征对于IDS意义重大。一方面可以减少数据的存储量，缓解模块之间数据通信及分析的压力；另一方面可以减少系统的训练时间，提高学习的准确性。因此，本文引入AdaBoost算法约减入侵数据的特征维数，使那些对分类影响甚微的特征被约减掉。

1.1 AdaBoost 算法简介

AdaBoost算法[2]是机器学习领域中的重要算法之一。该算法通过依次训练一组弱分类器，将它们集成为一个强分类器。基本过程是：每个训练样本被赋予一个权值，表明它被某个弱分类器选入训练集的概率。当一个弱分类器训练完成后，根据其在训练集上的分类结果对所有的样本权值进行调整。如果某个样本被当前弱分类器准确分类，那么它的权重就会被降低，则在构造下一个弱分类器的训练集时，它被选中的概率就被降低；相反，如果某个样本没有被正确分类，则它的权重就相应被提高，它入选下一个弱分类器训练集的概率被提升。通过这种方式， AdaBoost能够“聚焦于”那些比较困难(容易出现错分)的样本。集成后的强分类器的判决结果是所有弱分类器的判决结果的加权和。Schapire R. E.等人证明：使用AdaBoost算法能够得到既在训练集上具有低错误率又具备相当泛化能力的分类器[3]。

1.2 特征约减算法

AdaBoost 算法的目标是提高给定的学习算法的分类准确率。而本文从特征约减的角度，对该算法进行了修改，得

基金项目：广西自然科学基金资助项目(桂科基0575094)
作者简介：陶晓玲(1977－)，女，工程师、硕士研究生，主研方向：网络信息安全，网格技术；王 勇，教授、博士；罗 鹏，硕士
收稿日期：2007-12-17 **E-mail：**txl@guet.edu.cn

到算法基本思想：反复选择入侵特征构建两值弱分类器，选出一些分类能力最优的弱分类器，则挑选弱分类器的过程也就是选择特征的过程。训练过程中的每个弱分类器都是基于单特征的。每一轮循环挑选出在当前权重分布下误差最小的弱分类器，即选出具有最佳分类表现的相应特征，每个特征都是在已选出的特征所确定的权重信息前提下选出来的，最后得到的每一个弱分类器的权重在一定程度上还可以用来衡量不同特征对分类的贡献度，再综合采集成本消耗，约减部分特征。

具体算法描述如下：

(1)选取训练样本(x_1,y_1)，$(x_2,y_2),\cdots,(x_m,y_m)$，$x_i$表示每个入侵检测数据的特征向量，$y_i=\pm 1$ 为类别标识，其中$y_i=1$ 代表正常样本，$y_i=-1$ 代表异常样本。

(2)初始化权值，每个训练样本具有起始权值

$$w_{1,i}=\begin{cases}\dfrac{1}{2p} & y_i=1\\ \dfrac{1}{2q} & y_i=-1\end{cases}\quad(i=1,2,\cdots,m)$$

其中，p 和 q 分别表示正常样本和异常样本的数目。

(3) for $t=1,2,\cdots,T$

1)规一化权值 $w_{t,i}=\dfrac{w_{t,i}}{\sum_{i=1}^{m}w_{t,i}}$，使得$w_t$为一概率分布。

2)对于每一种未被选择的特征j，采用轮盘赌的方法选一训练子集训练弱分类器h_j(只限于单个特征)，分类误差计为 $\varepsilon_j=\sum_i w_{t,i}\left\{h_j(x_i)\neq y_i\right\}$。

3)选择具有最小误差的分类器h_j作为h_t，对特征j附加已选择标记(其在特征向量中的位置用l_t表示)。

4)更新权值 $w_{t+1,i}=w_{t,i}\beta_t^{1-e_i}$。当$x_i$被正确分类时，$e_i=0$；否则 $e_i=1$，并且 $\beta_t=\dfrac{\varepsilon_t}{1-\varepsilon_t}$。

5)t++。

(4)最终找出T个最优特征的分类器$h_1, h_2,\cdots,h_T$，其在特征向量中的位置是$l_1, l_2,\cdots,l_T$，每个分类器的权值为

$$\alpha_t=\ln\frac{1}{\beta_t}$$

2 入侵检测分类器的构造

根据上述特征约减算法得到约减的入侵特征后，就可以构造入侵检测分类器，以实现对入侵数据的分类判断。本文分别使用 Ada-加权、Ada-域值和支持向量机分类方法构造分类器。

2.1 Ada-加权分类器

Ada-加权分类器是按照传统的AdaBoost方法，直接根据特征约减和弱分类器的模型构造分类器。设未知的入侵检测数据为 x，选择 $l_1, l_2,\cdots,l_T$位置上的特征，定义分类函数为

$$f(x)=\sum_{t=1}^{T}\alpha_t h_t(x)$$

其中，h_t和α_t为 1.2 节算法中挑选出的弱分类器及相应的权值；T为约减后的特征个数；$|f(x)|$称为 x 的置信度，$|f(x)|$越大说明判断准确性越高。Ada-加权分类器的判决函数为

$$H(x)=\mathrm{sgn}(f(x))$$

其中，sgn() 为符号函数。若 $H(x)>0$，则判断 x 为正常数据；反之判断 x 为异常数据。

2.2 Ada-域值分类器

Ada-域值分类器是通过设定判决域值来识别分类结果的。与Ada-加权分类器相同的是：两者都是基于特征约减和弱分类器的模型构造的分类器。设未知的入侵检测数据的特征矢量为$\boldsymbol{x}$，选择$l_1, l_2,\cdots, l_T$位置上的特征，定义判决域值为 θ，令 θ 为所有弱分类器权值的平均值，则

$$\theta=\frac{1}{T}\sum_{t=1}^{T}\alpha_t$$

其中，α_t 为 1.2 节算法中挑选出的弱分类器对应的权值；T为约减后的特征个数。Ada-域值分类器的判决函数为

$$H(\boldsymbol{x})=\begin{cases}1 & \sum_{t=1}^{T}\alpha_t h_t(\boldsymbol{x})\geq\theta\\ -1 & \text{其他}\end{cases}$$

若 $H(\boldsymbol{x})=1$，则判断 $\boldsymbol{x}$ 为正常数据；$H(\boldsymbol{x})=-1$，则判断 $\boldsymbol{x}$ 为异常数据。

2.3 支持向量机分类器

SVM方法是从线性可分情况下的最优分类面发展而来的。它首先由Vaplink等人利用结构风险最小化(Structural Risk Minimization, SRM)原则提出，实际上是解决一个带不等式约束的二次凸规划问题[4]。对非线性问题，SVM首先通过非线性变换将输入空间变换到一个高维空间，然后在这个新空间中求得最优线性分类面，而这种非线性变换是通过定义适当内积核函数实现的。

本文采用的支持向量机分类器是将经过 AdaBoost 特征约减算法约减后的 T 个特征作为输入，使用 SVM 的 RBF 核函数为 $K(x,y)=\exp(-\gamma\|x-y\|^2)$。其中，$C$=1,$\gamma$=0.25，对入侵数据进行两类识别。

3 实验的建立及结果分析

3.1 实验环境

借鉴麻省理工学院林肯实验室在收集 IDS 的标准测试数据时提出的提供服务以及攻击可控制原则，本文的 Linux 入侵检测系统测试平台的构建原则为：(1)整个实验在正常的可控的网络环境中进行，服务器方提供 FTP, HTTP, SMTP, SSH/TELNET, SAMBA 等正常服务；(2)借助当前应用普遍的入侵检测系统以及防火墙，保证攻击的可控性和样本的性质。基于以上原则，设计实验环境如图 1 所示。

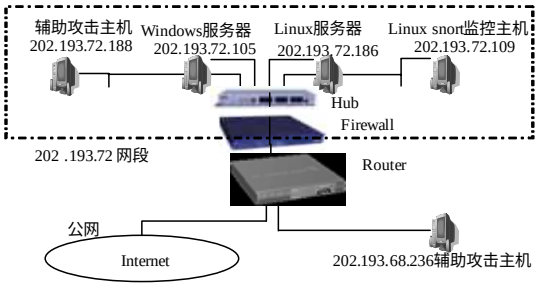

图 1 入侵检测的实验环境

在 Linux 服务器(202.193.72.186)上，安装了 Red Hat Linux 9 操作系统，在该服务器上分别开启了 Apache 以及 Tomcat 的 HTTP、Vsftpd 的 FTP、SSH、Samba、Sendmail 的 SMTP、MySQL 数据库等服务。对该服务器进行攻击并在其上进行入侵检测的数据采集工作。

3.2 网络攻击设计

图 1 的实验环境中 202.193.72.188 和 202.193.68.236 是辅助攻击的主机，当然，辅助攻击是在可控范围之内的。选择了多种有效的攻击工具，这些工具也是当前网络上黑客常用的或者是经典的工具：(1)综合扫描类工具：X-Scan, superscan, Retina, Hscan, netcat, bluescan, amap, nmap, 流光；

(2)拒绝服务攻击类工具：tfn2k, Jolt, udpflood, teardrop.c, Smurf, SYN Flood；(3)远程攻击工具：proft_put_down, httpd, Brutus；(4)普通授权用户提升到特权用户的攻击工具：hatorihanzo.c, Root.c；(5)后门程序：netcat, stcpshell, tcpbackdoor, adore, Q-2.4；(6)其他恶意脚本：sars。

3.3 特征提取

参考LIDS,snort等入侵检测系统的特征选取方法，本文的Linux入侵检测系统所选定的原始特征共 5 个类别、27 个参数，其中包含对入侵比较敏感的主机以及网络数据特征，如表 1 所示。而在特征提取的数据实际采集方面，本系统的设计主要针对设定的时间窗，对主机以及网络数据进行统计。时间窗的选择直接影响整个系统的检测率以及开销。本文经过理论计算和实验验证，最终将系统的时间窗设定为 1 s[5]。

表 1 Linux 入侵检测系统采集的特征

特征类别	特征含义(位置号)	取值类型
系统级特征	系统级 CPU 利用率(1)	实数
	用户级 CPU 利用率(2)	实数
	物理内存利用率(3)	实数
	SWAP 利用率(4)	实数
	敏感资源被占用的次数(11)	整数
	敏感系统调用次数(26)	整数
用户级特征	时间窗内用户错误登录次数(13)	整数
	登录用户的数目(14)	整数
	root 用户在登录用户中的比例(15)	实数
	敏感指令所占比例(16)	实数
进程级特征	运行进程的总数(9)	整数
	运行状态的进程所占比例(10)	实数
	敏感资源占用次数与占有敏感资源的进程数的比率(12)	实数
端口连接特征	TCP 开放端口数(5)	整数
	UDP 开放端口数(6)	整数
	RAW 开放端口数(7)	整数
	特殊端口连接数(8)	整数
	TCP 连接数(23)	整数
	UDP 连接数(24)	整数
	RAW 连接数(25)	整数
网络传输特征	SYN 包所占比例(17)	实数
	RST 包所占比例(18)	实数
	URG 包所占比例(19)	实数
	FLAG 包所占比例(20)	实数
	出站流量(21)	实数
	入站流量(22)	实数
	HTTP 传输错误数(27)	整数

3.4 实验结果分析

通过使用内核模块加载(LKM)、proc 编程等 Linux 高级编程技术，从 Linux 服务器中采集原始数据，并对原始数据进行预处理，形成样本数据。本文采用了 5 620 条数据作为实验数据。训练样本集采用 3 100 条样本数据，其中正常样本 1 600 条，异常样本 1 500 条。测试样本集采用 2 520 条样本数据，其中正常样本 1 479 条，异常样本 1 041 条。AdaBoost 特征约减算法及 3 种分类方法均在 PC 机(CPU P4 1.60 GHz，内存为 512 MB)上，利用 Visual C++ 6.0 开发工具实现。

用 AdaBoost 特征约减算法对 Linux 入侵检测系统所选定的 27 个原始特征进行了挑选和排序，结果为(重要性从大到小)：{23,17,1,5,4,22,15,19,2,3,18,14,6,7,8,9,10,11,26,12, 13, 20,16,21,24,25,27}。算法实现采用训练样本集数据，每轮的训练子集用轮盘赌方法选择 1 000 条样本数据，算法实现的时间为 653 s。根据得到的特征重要性排序序列，对入侵特征进行约减尝试。T 代表约减后的特征维数，分别取 T 为 27,25, 23,22,21,20，采用 Ada-加权分类器所得的识别结果见表 2。从表 2 可看出，T 为 22 时，识别效果最佳；如果采用 Ada-域值分类器，所得的识别结果见表 3。从表 3 可看出，T 为 21 时，识别效果最佳；如果采用支持向量机分类器，所得的识别结果见表 4。从表 4 可看出，T 为 22 时，识别效果最佳。结合数据采集的成本以及实际的需要，Linux 入侵检测系统约减掉原始特征中的{27,25,24,21,16,20}。由表 2~表 4 可看出，Ada-加权分类器和 Ada-域值分类器的识别效果相当，均好于支持向量机分类器；入侵特征选择得越多(指 19 个特征之上)，测试时间越长；而支持向量机分类器的测试时间要少于前两者分类器的测试时间，那是因为 Ada-加权和 Ada-域值分类器均要对多个分量分类器进行融合，造成时间上的消耗，不过还是可以达到实时性要求。

表 2 Ada-加权分类器分类结果

T	训练精度/(%)	测试精度/(%)	测试时间/s
27	100.00	93.42	8.592
25	100.00	93.61	8.031
23	99.81	93.61	7.380
22	99.24	93.40	7.040
21	98.97	92.97	6.759
20	99.81	92.21	6.158

表 3 Ada-域值分类器分类结果

T	训练精度/(%)	测试精度/(%)	测试时间/s
27	100.00	93.41	8.382
25	100.00	93.37	7.931
23	100.00	93.33	7.170
22	99.98	93.25	7.000
21	98.45	93.62	6.479
20	99.93	91.76	6.259

表 4 支持向量机分类器分类结果

T	训练精度/(%)	测试精度/(%)	测试时间/s
27	100.00	92.78	2.991
25	100.00	92.82	2.951
23	100.00	92.82	2.891
22	100.00	92.78	2.851
21	100.00	92.00	2.811
20	99.60	92.23	2.791

表 5 显示了经过特征约减后的 3 种分类器在测试集上的入侵检测性能比较。采用检测率(Detection Rate, DR)和虚警率(False Negative Rate, FNR) 2 个指标来考察 3 种分类方法的检测性能，计算公式如下：

检测率=分对的异常数据样本数/异常样本总数

虚警率=分错的正常样本数/正常样本总数

表 5 入侵检测性能比较(T=21)

分类器	检测率/(%)	虚警率/(%)
Ada-加权分类器	98.94	3.01
Ada-域值分类器	99.52	2.69
支持向量机分类器	98.08	4.22

相比之下，Ada-域值分类器的检测性能较好。如果对入侵特征不通过 AdaBoost 算法约减，而是随机地抽取 21 个特征，用支持向量机分类器进行分类，测试精度仅为 83.49%。结果说明：通过 AdaBoost 算法得到的特征比随机产生的特征所取得的分类效果好。

4 结束语

本文从特征选择和分类的角度，将改进的 AdaBoost 算法用于入侵特征的约减，并在此基础上构造 Ada-加权和 Ada-域值分类器，对支持向量机分类器与前两者进行分类结果对比。建立了一个模仿林肯实验室的 Linux 实时入侵检测实验环境，将特征约减算法和分类方法在其上进行验证，实验结果表明本文方法是有效的。 (下转第 206 页)