基于 Boosting 算法的入侵检测

况 夯
KUANG Hang

重庆教育学院,重庆 400067
Chongqing Education College,Chongqing 400067,China

KUANG Hang.Intrusion detection based on boosting method.Computer Engineering and Applications,2008,44(4):151-154.

Abstract: A novel method is proposed for intrusion detection based on improved boosting BP neural network.In order to improve the precision of the BP neural network for intrusion detection,the improved boosting algorithm is used to build an integration-neural network.The improved boosting adopts a new method to acquire parameters; the weighted parameters of weak classifiers are determined not only by the error rates,but also by their abilities to recognize the positive samples.Simulated experiments with KDD Cup 1999 network connections data which have been preprocessed with methods of features selection and normalization have shown that the proposed is effective for intrusion detection owing to excellent performance of the higher attack detection rate with lower false positive rate.

Key words: intrusion detection;Boosting method;BP neural network

摘 要:提出了一种新颖的基于 boosting BP 神经网络的入侵检测方法。为了提高 BP 神经网络的泛化能力,采用改进的 Boosting 方法,进行网络集成。Boosting 方法采用更有效的参数求解方法,即弱分类器的加权参数不但与错误率有关,还与其对正样本的识别能力有关。对"KDD Cup 1999 Data"网络连接数据集进行特征选择和归一化处理之后用于训练神经网络并仿真实验,得到了较高的检测率和较低的误报率,仿真结果表明,提出的入侵检测方法是有效的。

关键词:入侵检测;Boosting 方法;BP 神经网络
文章编号:1002-8331(2008)04-0151-04 **文献标识码**:A **中图分类号**:TP393

1 引言

随着互联网络的飞速发展,网络安全正日益得到人们的关注。近年来,使用互联网进行电子商务和在线消费的活动越来越频繁,共享网络计算机系统的安全问题越来越突出。入侵检测系统应运而生,它通过对计算机网络或系统中若干关键点的信息进行收集分析,从中发现是否有违反安全策略的行为和被攻击的迹象,是一种集检测、记录、报警、响应为一体的动态安全技术[1]。

入侵检测系统的功能是检测出入侵事件的发生,入侵检测可以看作是一个分类问题,把给定的审计数据分为正常数据和异常数据。入侵方式不同,入侵检测的策略和模型也不一样。目前已有统计方法[2]、专家系统方法[3]、神经网络方法[4]等,它们有各自的优缺点。统计方法依赖于一些假设,如审计数据或用户行为)的分布符合高斯分布,实际上,用户行为具有随机性,这种假设可能导致较高的误警率。基于专家系统的入侵检测很难包括未知的进攻行为,另外还有效率问题。神经网络方法是很有潜力的方法,神经网络可以利用大量实例通过训练的方法学会掌握知识,获得预测能力,还可以向神经网络展示新发现的入侵攻击实例,通过再训练可以使神经网络的攻击模式产生反应,从而使入侵检测系统具有自适应能力。

本文提出了一种新颖的基于 Boosting BP 神经网络的入侵检测方法。在入侵检测系统中使用神经网络,对用户连接行为特征进行检测和识别,即用不同用户连接行为的历史数据作为输入,为了进一步提高神经网络的泛化能力,采用 Boosting 方法[5],进行网络集成。Boosting 方法采用更有效的参数求解方法,即弱分类器的加权参数不但与错误率有关,还与其对正样本的识别能力有关。所设计的分类器以加权投票方式进行分类决策。本文的结构如下:第 2 章提出改进的 boosting 算法,第 3 章确定 BP 神经网络的结构, 提出 BP 神经网络的训练算法和基于 boosting BP 神经网络的入侵检测算法, 并给出了具体的算法步骤,第 4 章为仿真实验与结果分析,第 5 章是结论。

2 改进的 boosting 算法

Boosting[5]方法的基本思想是:给定一弱学习算法和一训练集$(x_1,y_1),(x_2,y_2),\cdots,(x_n,y_n)$,这里 x_i 为第 i 个训练样本的输入,y_i 为分类问题的类别标志。算法开始给每一个训练样本赋予相等的权值 $1/n$,然后用该学习算法对训练集样本训练 T 轮,每次训练后,对训练样本赋予权值时同时考虑对正样本的识别能力,在错误率 E_t 相同的情况下,那些对正样本识别能力更强的弱分类器具有更大的权值, 从而得到一个预测函数序列 h_1,

作者简介:况夯(1974-),男,硕士,讲师,主要研究方向:模式识别,网络安全,电子商务。
收稿日期:2007-07-03 **修回日期**:2007-08-20

$h_2,\cdots,h_T$,其中每个 h_j 也对应一定的权值,预测效果好的预测函数的权值较大,反之较小。最终的预测函数 H 采用加权投票方式对新样本进行判别。改进的 Boosting 算法具体描述如下:

(1)输入:$(x_1,y_1),(x_2,y_2),\cdots,(x_n,y_n);x_i\in X,y_i\in Y=\{-1,+1\}$

初始化:$D_1(i)=1/n$

(2)For $t=1,\cdots,T$

a.在 D_t 下训练,得到预测函数 h_t

b.计算该预测函数的错误率

$$E_t=\sum_{h_t(x_i)\neq y_i}D_t(i)$$

以及识别正确的正样本的权值和

$$p_t=\sum_{y_i=1,h_t(x_i)=1}D_t(i)$$

c.求解弱分类器 h_t 的加权系数

$$\alpha_t=\frac{1}{2}\ln\left(\frac{1-E_t}{E_t}\right)+k\cdot e^{p_t}$$

K 为一个常数,其取值满足在本次循环中,令最小错误率的上界下降。

d.根据错误率更新样本的权值:

$$D_{t+1}(i)=\frac{D_t(i)}{Z_t}\times\begin{cases}e^{-\alpha_t},h_t(x_i)=y_i\\e^{\alpha_t},h_t(x_i)\neq y_i\end{cases}=\frac{D_t(i)e^{(-\alpha_t y_i h_t(x_i))}}{Z_t}$$

其中,Z_t 是归一化因子,即

$$Z_t=\sum_i D_t(i)e^{(-\alpha_t y_i h_t(x_i))}$$

(3)对于未知样本,输出:$H(x)=sign(\sum a_t h_t(x))$

新的算法采用了与传统 Boosting 算法不同的加权参数的求解公式:

$$\alpha_t=\frac{1}{2}\ln\left(\frac{1-E_t}{E_t}\right)+k\cdot e^{p_t}$$

其中,k 为常数,$p_t=\sum_{y_i=1,h_t(x_i)=1}D_t(i)$。

显然,p_t 是第 t 次循环的弱分类器中所有被识别正确的正样本的权值和,它能够代表弱分类器 h_t 对正样本的识别能力,而 $k\cdot e^{p_t}$ 是 p_t 的增函数,这样求得的新的加权参数,在错误率 E_t 相同的情况下,那些对正样本识别能力更强的弱分类器具有更大的权值。

3　基于 Boosting BP 神经网络的入侵检测

3.1　BP 神经网络

误差反向传播算法(Error Back Propagation),简称 BP 算法。采用 BP 算法的多层神经网络模型一般称为 BP 网络,它是目前人工神经网络中研究最深入、应用最广泛的一类网络。BP 网络由输入层、中间层和输出层组成,中间层也就是隐含层可以是一层或多层。本文采用的是 3 层 BP 网络,即隐含层只有一层,其结构如图 1 所示。

本文使用的 BP 网络是用作分类器,其类别数为 2(即正常或异常),所以输出层的节点数为 2。选择用户连接行为的历史数据的 18 个特征作为输入,隐层节点数的选择是一个十分复杂的问题。如果数目过少,则网络不"强壮",不能识别以前没有看过的样本,容错性差;但如果数目过多,就会使学习时间过长,网络的泛化能力降低,而且误差也不一定最佳,因此存在一个最佳的隐层节点数。先根据经验公式:$n_1=\sqrt{n+m}+\alpha$ 和 $n_1=\mathrm{lb}\ n$(其中 n 为输入神经元数,m 为输出神经元数,α 为 1~10 之间的常数)计算出隐层节点数的取值范围(6~15),然后根据这个范围做相关的实验,改变隐层节点数,比较在训练集相同的情况下网络的收敛速度和网络分类的正确率。实验结果表明:如果隐层节点数在 8~13 之间变化,网络分类的正确率比较高,但收敛速度不同。综合考虑各个因素,选择的隐层节点数是9。

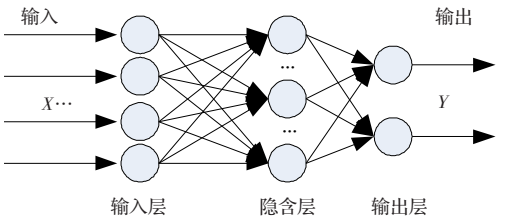

图 1　BP 神经网络结构

3.2　训练过程

假设输入层、中间层和输出层的单元数分别是 N、L 和 M。$\boldsymbol{X}=(x_0,x_1,\cdots,x_{N-1})$ 是 BP 网络的输入矢量,$\boldsymbol{H}=(h_0,h_1,\cdots,h_{L-1})$ 是中间层输出矢量,$\boldsymbol{Y}=(y_0,y_1,\cdots,y_{M-1})$ 是网络的实际输出矢量,并且用 $\boldsymbol{D}=(d_0,d_1,\cdots,d_{M-1})$ 来表示训练组中各样本的目标输出矢量。输入单元 i 到隐单元 j 的权重是 V_{ij},而隐单元 j 到输出单元 k 的权重是 W_{jk}。另外用 θ_k 和 ϕ_j 来分别输出单元和隐单元的阈值。于是中间层各单元的输出为

$$h_j=f(\sum_{i=0}^{N-1}V_{ij}x_i+\phi_j)\tag{1}$$

而输出层各单元的输出是:

$$y_k=f(\sum_{i=0}^{L-1}W_{jk}h_j+\theta_k)\tag{2}$$

其中 $f(*)$ 是激励函数,这里采用的是 S 型函数:$f(x)=\frac{1}{1+e^{-x}}$。

在上述条件下,网络的训练过程如下:

(1)选定训练组,从正常和异常的样本集中分别随机地选取 500 个样本作为训练组。

(2)将各权重 V_{ij}、W_{jk} 和阈值 ϕ_j、θ_k 设置成小的接近于 0 的随机值,并初始化精度控制参数 ε 和学习率 α。

(3)从训练组中取一个输入样本 X 加到网络,并给定它的目标输出矢量 $\boldsymbol{D}$。

(4)利用式(1)计算出一个中间层输出矢量 $\boldsymbol{H}$,再用式(2)计算出网络的实际输出矢量 $\boldsymbol{Y}$。

(5)将输出矢量中的元素 y_k 与目标矢量中的元素 d_k 进行比较,计算出 M 个输出误差项:

$$\delta_k=(d_k-y_k)y_k(1-y_k)$$

对中间层的隐单元也计算出 L 个误差项:

$$\delta_j^*=h_j(1-y_j)\sum_{k=0}^{M-1}\delta_k W_{jk}$$

(6)依次计算出各权重和阈值的调整量:

$$\Delta W_{jk}(n)=(\alpha/(1+L))*(\Delta W_{jk}(n-1)+1)*\delta_k*h_j\tag{3}$$

$$\Delta V_{ij}(n)=(\alpha/(1+N))*(\Delta V_{ij}(n-1)+1)*\delta_j^**x_i\tag{4}$$

$$\Delta\theta_k(n)=(\alpha/(1+L))*(\Delta\theta_k(n-1)+1)*\delta_k\tag{5}$$

$$\Delta\phi_j(n)=(\alpha/(1+L))*(\Delta\phi_j(n-1)+1)*\delta_j^*\tag{6}$$

(7)调整权重:$W_{jk}(n+1)=W_{jk}(n)+\Delta W_{jk}(n)$,$V_{ij}(n+1)=V_{ij}(n)+\Delta V_{ij}(n)$和阈值:$\theta_k(n+1)=\theta_k(n)+\Delta\theta_k(n)$,$\phi_j(n+1)=\phi_j(n)+\Delta\phi_j(n)$。

(8)当 k 每经历 1~M 后,判断指标是否满足精度要求:$E\leqslant\varepsilon$,其中 E 是总误差函数,且 $E=\frac{1}{2}\sum_{k=0}^{M-1}(d_k-y_k)^2$。如果不满足,就返回(3),继续迭代。如果满足,就进入下一步。

(9)训练结束,将权重和阈值保存在文件中。这时可以认为各个权重已经达到稳定,分类器形成。再一次进行训练时,直接从文件导出权重和阈值进行训练,不需要进行初始化。

3.3　基于 boosting BP 神经网络的入侵检测

入侵检测就是根据用户输入的历史信息判断网络是否正常。由于神经网络存在着结构不固定,分类精度不高等问题,针对这种情况,通过 boosting 方法对 BP 神经网络进行加强,得到一个使用弱分类算法但同时具有强分类性能的分类器,可以很好地提高事件检测的检测率和降低误报率。具体算法步骤如下:

(1)初始化样本权重:$D_1(i)=\frac{1}{n}$

//表示第一次迭代,训练样本权重为$\frac{1}{n}$

(2)对于 t=1 to T 进行迭代　　//T 为迭代次数

在 $D_t(i)$下训练,使用 BP 神经网络训练得到弱分类假设:

$h_t:X\to\{+1,-1\}$;　　//得到第次预测函数

a.计算 h_t 的错误率:$E_t=\sum_{h_t(x_i)\neq y_i}D_t(i)$　//错分样本

b.计算分类假设 h_t 权值:$\alpha_t=\frac{1}{2}\ln\left(\frac{1-E_t}{E_t}\right)+k\cdot e^{p_t}$

//h_t 的权重

c.更新权值:$D_{t+1}(i)=\frac{D_t(i)}{Z_t}\times F(E_t)=\frac{D_t(i)}{Z_t}\times\begin{cases}e^{-\alpha_t},h_t(x_i)=Y_i\\e^{\alpha_t},h_t(x_i)\neq Y_i\end{cases}$

//Z_t为归一化因子

(3)最后的输出分类函数:

$$H(x)=\arg\max_{y\in Y}\left(\sum_{t:h_t(x)=y}\alpha_t\right)$$

4　仿真实验与结果分析

4.1　数据预处理

用于本文实验的数据集来源于“KDD Cup 1999 Data”[6],该数据集中的每个连接都带有一个或标有正常(Normal)、或者具体攻击类型的类标识并且所有的这些数据无一例外的都有 41 个特征属性。这些特征中很多属于冗余特征,因此只选择其中 18 个能够体现用户行为的部分特征作为研究对象,即 duration(连接持续的时间)、protocol_type(协议类型)、service(目标网络服务)、src_bytes(源地址到目标地址的字节数)、dst_bytes(目标地址到源地址的字节数)、flag(网络连接状态,正常或错误)等。

由特征提取模块获得的网络连接记录信息格式复杂,必须将这些信息转换成 BP 神经网络能够处理的向量形式。对于数据的向量化处理,遵循文献[7]提出的规则。第 1 步,先将所有类型的数据转换成以二进制表示的数字形式。转换时采用文献[7]提出的基于距离度量函数 HVDM 的方法,对数据进行归一化处理。例如:对于协议类型的特征值{tcp,udp,icmp},则转换成二进制分别为{1,0,0}、{0,1,0}、{0,0,1}。第 2 步,对这些特征值的范围进行处理,使得每类特征数据的取值范围在区间[0,1]中。这样处理一方面可以避免取值范围大的特征支配那些取值范围小的特征;另一方面可以降低机器的计算时间。

4.2　实验结果分析

由于从训练集中得到的最终的归类已经作了正常类或异常类的标识,因此我们可以应用该分类器进行入侵检测,在进行实时检测时,首先对每一个连接数据进行解析,即只需要取出最能体现该数据连接特征的 18 个数据属性,然后对数据作归一化数据处理,将其归一到[0,1]的特征空间中以后,作为神经网络的输入,将其输入到已经训练好的神经网络输入层,根据其输出来识别其为正常类或为异常类。

入侵检测算法的好坏主要取决于如下两个指标,即:检测率 ADR(Attack Detection Rate)和误报率 FPR(False positive Rate)。所谓检测率就是被检测到的攻击数目占总攻击数的百分比,而误报率则是正常的连接被误报为入侵连接占正常连接总数的百分比。如果运用入侵检测算法得到的检测率很高,也就是漏报率很低,同时误报率也很低则说明几乎所有的入侵攻击都可以被系统检测到,并且很少出现误报的现象,则证明该算法的性能是优良的。

下面我们给出了 3 个实验的结果:(1)候选样本数量的选择;(2)boosting 算法的迭代次数对比;(3)不同分类方法的选择。

实验 1　候选样本数量的选择。

本实验主要研究候选样本数量对检测结果的影响。候选样本集中正负样本比例固定为 1:1,改变候选样本集的规模,如表 1 所示;测试集中样本数为 1 000,样本组成与候选样本集相同,boosting 算法的迭代次数为20,表 2 为实验 1 结果。

表 1　候选样本集的组成

U 中样本总数	正常样本与异常样本比
200	1:1
500	1:1
1 000	1:1

表 2　实验 2 的结果

U 中样本总数	检测率/%	误报率/%	漏报率/%
200	97.34	0.37	0.64
500	99.46	0.28	0.56
1 000	99.87	0.20	0.42

从表 2 可以看出,当候选样本集中样本数越多,分类器的检测精度越高,这是因为样本数量越多,代表正常和异常之间的区别的数据越多,从而使得 BP 神经网络的检测精度随着学习样本的增多而不断提高。

实验 2　boosting 算法的迭代次数对入侵检测性能的影响。

在这个实验中,我们在样本总数为 1 000,正常样本与异常样本比为 1:1 的条件下进行实验,分别对 boosting 算法的迭代次数为 5 次,10 次,15 次,20 次进行实验,发现当迭代次数增加时,检测的性能也增加。实验结果如表 3 所示。

表 3　不同迭代次数的比较结果

迭代次数	检测率/%	误报率/%	漏报率/%	平均检测时间/s
5	75.43	2.01	2.97	0.38
10	84.32	1.52	2.42	0.40
15	93.64	0.63	0.86	0.42
20	99.87	0.20	0.42	0.44

实验 3 不同分类方法的选择。

在这个实验中,在样本总数为 1 000,正常样本与异常样本比为 1:1 的条件下比较 3 种不同的分类方法:单独的 BP 神经网络方法,基于 Boosting BP 神经网络和改进的 Boosting BP 神经网络方法。根据实验 2,boosting 算法的迭代次数越多,检测性能越好,但是检测时间也随之增加,由于迭代 20 次检测率已经达到 99.87%,所以在本实验中,迭代次数为 20 次,实验结果如表 4 所示。

表 4 不同分类方法的比较结果

不同的分类方法	检测率/%	误报率/%	漏报率/%	平均检测时间/s
BP 神经网络	72.36	2.36	3.24	0.37
Boosting BP 神经网络	88.21	0.97	1.67	0.44
改进的 boosting BP 神经网络	99.87	0.20	0.42	0.44

从表 4 中可以看出,使用 boosting 算法对神经网络进行加强,可以提高检测率,降低误报率和漏报率,改进的 boosting 算法进一步提高了入侵检测的性能,而且使用 boosting 算法的时间仅比神经网络多 0.07 s。

5 结论

入侵检测是非常重要的研究课题,本文将入侵检测问题转化一个分类问题,提出了一种新颖的基于 Boosting BP 神经网络的入侵检测方法。以不同用户连接行为的历史数据作为 BP 神经网络分类器的输入,进行入侵检测;为了提高神经网络的泛化能力和入侵检测的准确性,采用改进的 Boosting 方法对 BP 神经网络进行集成,分类器以加权投票方式进行分类决策。实验结果证明改进的 boosting 算法比传统算法性能有了较大的改进,采用本文提出的算法进行入侵检测,取得了较好的结果。

参考文献:

[1] 马传香,李庆华,王卉.入侵检测研究综述[J].计算机工程,2005,31(3):4-6.

[2] Bykova M,Ostermann S,Tjaden B.Detecting network intrusions via a statistical analysis of network packet characteristics[C]//Proceedings of the 33rd Southeastern Symposium in System Theory,2001.

[3] Jackson K,DuBois D,Stalling C.An expert system application for network.intrusion detection[C]//Proceedings of the 14th National Computer Security Conference,1991-10:215-225.

[4] Bonifacio J M.Neural networks applied in intrusion detection systems[C]//Proc of the IEEE World Congress on Comp Intell(WCCI98),1998.

[5] Freund Y,Schapire R E.A decision-theoretic generalization of on-Line learning and an application to boosting[J].Journal of Computer and System Sciences,1997,55(1):119-139.

[6] KDD Cup 1999 Data[EB/OL].[1999].http://kdd.ics.uci.edu /databases/kddcup.html.

[7] Wilson D R,Tony R M.Improved heterogeneous distance functions[J].Journal of Artificial Intelligence Research,1997,6(1):1-34.

(上接 115 页)

密钥报文进行过滤

$mcastmonitor print-trace

其仿真的结果如图 2 和图 3 所示。

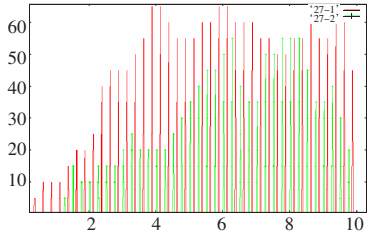

图 2 成员数为 27,两种方案的密钥数量变化示意图

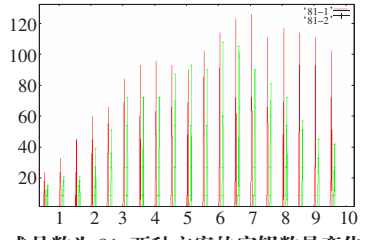

图 3 成员数为 81,两种方案的密钥数量变化示意图

图中,'27-1'曲线和'81-1'曲线表示文献[1]中提出方案的密钥报文变化曲线,'27-2'曲线和'81-2'曲线表示文中提出方案的密钥报文变化曲线。从图中可以看出,文中提出的优化方案要优于文献[1]中提出的方案;从曲线的走势来看,由于是批量密钥更新,所以在更新周期到达时才产生密钥,因此密钥报文呈突发性。

由此可见,优化的批处理将密钥更新主要集中在高频变动子树中,从而进一步加大路径的重叠数,减少加密计算量和密钥分配的通信量。

5 结论

论文对文献[1]中提出的批密钥更新优化方案进行了一些修改,提出建立动态变动子树,限制成员加入的位置,增大批量密钥更新过程中的路径重叠概率,降低加密次数,解决了基于 LKH 树的密钥更新方案在动态多播环境下的扩展问题。

参考文献:

[1] 韩秀林,王行愚.基于 LKH 树的动态多播群组批密钥更新方案的优化设计[J].计算机工程与科学,2005,27(8):20-23.

[2] Li X S,Yang Y R,Gouda M G,et al.Batch rekeying for secure group communications[C]//The 10th Int'l World Wide Web Conference,HongKong,2001.

[3] Rafaeli S,Mathy L,Hutchison D.LKH+2:an improvement on the LKH+ algorithm for removal operations[EB/OL].Internet draft(work inprogress),Internet Eng,Task Force,http://www.watersprings.org/pub/id/draft-rafaeli-lkh2-00.txt,2002.

[4] Pegueroles J,Rico-Novella F,Hernández-Serrano J,et al.Improved LKH for batch rekeying in multicast groups[C]//Proc of the IEEE Int'l Conf on Information Technology Research and Education(ITRE 2003).New Jersey:IEEE Press,2003:269-273.

[5] Pegueroles J,Rico-Novella F.Balanced batch LKH:new proposal,implementation and performance evaluation[C]//IEEE Symposium on Computers and Communications-ISCC'2003,2003.

[6] 许勇,陈恺.安全多播中基于成员行为的 LKH 方法[J].软件学报,2005,16(4):601-608.

[7] Fall K,Varadhan K.The NS Manual(formerly NS Notes and Documentation)[EB/OL].http://www.isi.edu/nsnam/ns/ns-documentation.html,2003.