基于集成学习的入侵检测方法

徐 冲 王汝传 任勋益
(南京邮电大学计算机学院 南京 210003)

摘 要 为解决传统入侵检测中存在的检测效率低、对未知的入侵行为检测困难等问题,提出了将改进的BP神经网络算法和支持向量机集成的入侵检测模型。实验表明,集成改进的BP神经网络和支持向量机与检出率最好的单个神经网络、单个SVM相比检测率有所提高,同时提高了对未知入侵行为的识别。

关键词 入侵检测,集成学习,BP神经网络,支持向量机
中图法分类号 TP393 **文献标识码** A

Ensemble Learning Based Intrusion Detection Method

XU Chong WANG Ru-chuan REN Xun-yi
(College of Computer, Nanjing University of Posts and Telecommunications, Nanjing 210003, China)

Abstract In order to solve the problem of low detection rate for novel attacks and the difficulties in detecting unknown intrusions existing in traditional intrusion systems, the paper proposed a model based on ensemble learning in improved BP neural networks and support vector machines. Experiments show that using the ensemble learning method, the detection rate is higher than that of using any individual networks and svm. So it has a better detection rate not only to the known intrusion, but also to the unknown intrusion.

Keywords Intrusion detection, Ensemble learning, Back propagation neural network, Support vector machine

网络技术和网络规模的不断发展使得网络安全成为全球性的重要问题之一。迅速、有效地发现各类新的入侵行为,对于保证网络系统的安全十分重要。入侵检测是一种通过监视网络系统的运行状态,进而发现各种攻击行为的信息安全技术。入侵检测领域中研究的重点之一是分类算法,常用的入侵检测技术有概率统计、神经网络、支持向量机、人工免疫等。但单一的分类算法由于自身的原因,总存在各种缺陷,如神经网络算法的泛化能力差、收敛速度慢等。本文在研究两种改进的传统网络入侵检测学习器——BP神经网络和SVM基础上,提出了改进的BP神经网络和SVM集成学习的入侵检测模型,以提高入侵检测的检测率和对新型攻击的识别。

1 人工神经网络与支持向量机

1.1 人工神经网络(Artificial Neural Network, ANN)

近年来,基于误差反向传播算法(Error Back2 Propagation,BP算法)的多层前馈神经网络正广泛地被应用于模式识别、信号处理与图像处理、人工智能及控制等领域。然而在实际应用中,BP算法存在一些不足,主要是收敛速度很慢、往往收敛于局部极小点、数值稳定性差等问题,很难满足实时入侵检测的要求。为此本文引入3种算法来改进BP神经网络的性能。

1.1.1 弹性BP算法(Resilient back propagation, RP)

BP神经网络通常采用S型激励函数的隐含层(S型函数见式(1))。S型函数常被称为"压扁"函数,它将一个无限的输入范围压缩到一个有限的输出范围内,导致算法中的梯度幅值很小,可能得其对网络权值的修正过程几乎停滞下来。

$$f(x)=\frac{1}{1+\exp(-ax)},a>0;-\infty<x<\infty \quad (1)$$

针对这一问题,弹性BP算法只取偏导数的符号,不考虑偏导数的幅值。权值更新的方向由偏导数的方向决定,而权值变化的大小则由一个独立的"更新值"确定。若在两次连续的迭代中,目标函数对某个权值的偏导数的符号不变,则增大相应的"更新值";反之,则减小相应的"更新值"。

1.1.2 尺度化共轭梯度反向传播算法(Scaled conjugate gradient algorithm, SCG)

BP算法是最小均方误差的近似最速下降算法。从理论上讲,当步长取任意小时,该方法一定会使误差趋向局部或全局极小。但是当步长过小时,收敛速度很慢,使得网络的学习失去意义。SCG算法能在一定程度上克服最速下降迭代路径呈锯齿形现象的缺点,不必计算或存储二阶导数信息就具有二次终止性,因此在较大的规模问题中十分有用。在各种优化算法中,选定不同的搜索方向和步长,对算法的收敛性至关重要。BP算法选择的搜索方向是沿梯度的负方向搜索,而

到稿日期:2009-08-06 返修日期:2009-11-09 本文受国家自然科学基金(60973139,60773041),江苏省自然科学基金(BK2008451),国家高科技863项目(2007AA01Z404,2007AA01Z478),现代通信国家重点实验室基金(9140C1105040805),国家和江苏省博士后基金(0801019C,20090451240,20090451241),江苏高校科技创新计划项目(CX08B-085Z,CX08B-086Z)和江苏省六大高峰人才项目(2008118)资助。

徐 冲(1986-),男,硕士生,主要研究方向为网格计算、分布式系统和计算机软件在通信中的应用;王汝传(1943-),男,教授,博士生导师,主要研究方向为计算机软件、计算机网络和网格、对等计算、信息安全、无线传感器网络、移动代理和虚拟现实技术等,E-mail:wangrc@njupt.edu.cn。

SCG 算法则利用一维搜索所得到的极小点处的最速下降方向生成共扼方向作为搜索方向，比最速下降法在速度和效果上有很大的改进。设梯度向量为 g，共扼向量为 p，则第 kK 次的共扼方向为：

$$p(k)=\begin{cases}-g(k), k=0\\ -g(k)+B_k p(k-1), k\geqslant 1\end{cases} \tag{2}$$

式中，$\beta_k=(g(k)-g(k-1))^T g(k)/(g(k-1)^T g(k-1))$。权值的修正公式为 $w(k+1)=w(k)+\eta * p(k)$。

1.1.3 Levenberg-Marquard 算法(LM)

LM 算法使用雅可比矩阵(Jacobin matrix)计算，而标准 BP 算法使用海森矩阵(Hessian)计算。海森矩阵由 w_{ij} 的二阶偏导数构成，正确计算很困难，当精度不高时，难于达到最优点；在维数很高时，计算海森矩阵的逆矩阵需要消耗较多的时间，限制了该方法的实际应用。雅可比矩阵由网络误差相对于权值和偏差的一阶导数构成，是海森矩阵的近似矩阵，可以通过标准的方向传播方法来计算。当误差函数具有平方和的形式时，得到的连接权调整量为 $\Delta w_{ij}=[J^T J+\mu I]^{-1} J^T E_i$，其中 J 为雅可比矩阵，E_i 是网络误差向量，I 是单位矩阵。当 μ 为 0 时，上式是使用近似海森矩阵的牛顿法；当 μ 较大时，上式变成了具有较小步长的梯度下降法。此时海森矩阵可以近似为 $H=J^T J$，梯度为 $g=J^T E_i$。

1.2 支持向量机(Support Vector Machines,SVM)

SVM 的主要思想是建立一个最优决策超平面，使得该平面两侧距平面最近的两类样本之间的距离最大化，如图 1 所示。SVM 最初是为两类分类问题而设计的。

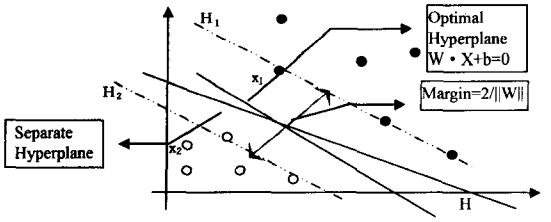

图 1 SVM 最优超平面

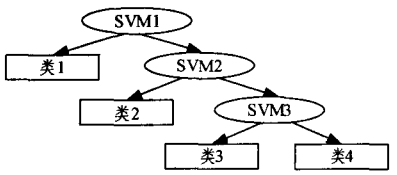

图 2 基于二叉树的 SVM 分类

而在实际应用中，多类分类问题更为普遍，如何将 SVM 的优良性能推广到多类分类中去，已成为 SVM 研究的一个热点问题。本文采用基于二叉树的多类 SVM 分类方法。对于 K 类的训练样本，训练 $K-1$ 个 SVM。第 1 个 SVM 将第一类样本作为正的训练样本，第 2,3,…,K 类训练样本作为负的训练样本训练 SVM1，第 i 个 SVM 将 i 类样本作为正的训练样本，第 $i+1,i+2,\cdots,K$ 类训练样本作为负的训练样本训练 SVMi，直到第 $K-1$ 类样本作为正样本，以第 K 类样本为负样本训练 SVM($K-1$)。图 2 给出的是一个拥有 4 类训练样本的基于二叉树的多类 SVM 分类示意图。

2 集成学习

2.1 集成学习

集成学习(Ensemble Learning)的主要思想是利用多个分类器来解决同一个问题，目的是更有效地提高学习系统的泛化能力。集成学习通常分为两个步骤：首先，采用单个学习算法对样本分别进行训练；然后，对单个网络的输出按某种方法进行集成，得到最后的结果。集成网络常用的方法包括 Bagging 和 Boosting 等。网络的集成输出中，分类问题常采用相对多数和绝对多数法，回归问题常采用加权平均和简单平均法。

2.2 人工神经网络和支持向量机集成

研究表明，当集成学习中各子分类器差异度较大时，才有较好的分类正确性。首先构建 4 个不同的子分类器，即弹性 BP 神经网络、SCG-BP 神经网络、LM-BP 神经网络和 SVM。如图 3 所示，对数据集进行预处理，包括对数据包的数值化、归一化和特征选取后，4 个子分类器经过训练学习对预处理好的数据进行分析，当结果不一致时，采用多数投票的方法判断该数据包的属类。

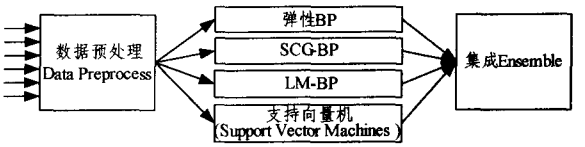

图 3 集成人工神经网络和支持向量机结构图

3 实验与分析

3.1 实验数据

为考察入侵检测的性能，采用 KDDCUP99 数据集，在 Matlab 平台上进行了实验研究。KDDCUP99 数据集中包含大量的正常网络流量数据和各种入侵行为数据，它们具有较好的代表性。数据集中共包含 38 种入侵行为，属于以下 4 种范畴：DoS (Denial-of-Service)：拒绝服务攻击，如 SYN FLOOD,land;Probe:非法监听和探测，如各种端口扫描和漏洞扫描;R2L(Remote-to-Local)：远程权限获取，如基于字典的口令猜测;U2R(User-to-Root)：本地用户非法提升权限的攻击，如缓冲区溢出攻击等。

在 KDDCUP 测试集，我们将整个数据集分为 6 类：正常数据、DoS、Probe、R2L、U2R，再将 4 类攻击中的每一类攻击类型分成两个部分：已知攻击和未知攻击。其中，已知攻击表示曾经在训练集中标记过的攻击，未知攻击指在训练集中没有标记过的攻击类型。把每一类攻击中的未知部分全归为“其他”一类，这样一共就有 6 类。从 KDDCUP 中分别随机抽取 10000 条记录作为训练集和 30000 条记录作为测试集，表 1 中罗列了每一类攻击中所包括的攻击类型以及样本总数。在对训练集和测试集进行数值化和归一化预处理后，利用 Matlab 提供的工具箱函数，在相同训练集和测试集的情况下分别利用 ANN、SVM 以及集成 ANN 和 SVM 进行检测分析。

表 1 训练集和测试集的分类情况

分类	具体攻击类型	训练样本-10000 条记录	测试样本-30000 条记录
正常	nomal	1875	6004
DoS	SYN Flood(Neptune),Ping of Death(P. D),Land 和 smurf	7944	23481

Probe	Ipsweep, satan, saint, nmap 和 Portsweep	83	233
R2L	IP spoofing, WareZclient, Ftp_write, Multihop	20	72
U2R	Buffer overflow	0	3
其他	unknown intrusion	78	207

3.2 单个 BP 神经网络的检测

设置人工神经网络的性能目标是 0.01，SCG-BP 用了 490 步训练、LM-BP 用了 500 步训练、弹性 BP 用了 499 步训练分别达到训练目标。图 4、图 5 和图 6 分别显示了经过训练后的 SCG-BP、LM-BP 和弹性 BP 对测试集数据进行预测的误检个数和误检率，以及期望的输出和实际输出。图中纵轴的 0 代表正常数据，1 代表 DoS 攻击，2 代表 Probing 攻击，3 代表 R2L 攻击，4 代表 U2L 攻击，5 代表未知攻击；横轴表示 30000 条测试数据。对于 30000 条记录的测试样本，SCG 有 813 个误检，误检率为 2.71%；弹性 BP 有 998 个误检，误检率为 3.3267%；LM-BP 有 1087 个误检，误检率为 3.6233%。在用 3 种改进的 BP 算法检测时，我们发现 LM-BP 算法训练时间比较长，而且总体预测率相对较低，但能识别样本数量少的攻击类型和未知攻击；而弹性 BP 在对权值和阈值更新时只考虑梯度的符号，从而提高了学习效率。SCG 算法结合了 LM 算法中的模型置信区间方法和共轭梯度算法，避免了耗时巨大的线搜索过程，从而提高了网络的训练速度。

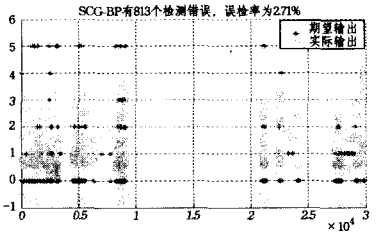

图 4 SCG-BP 误检个数和误检率

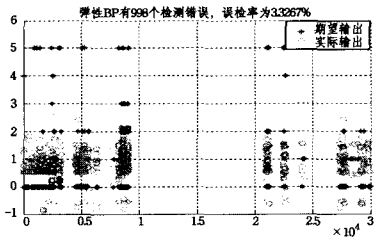

图 5 弹性 BP 误检个数和误检率

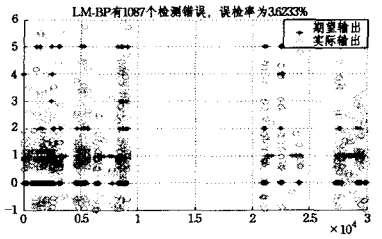

图 6 LM-BP 误检个数和误检率

从仿真结果看，SCG 的误检率最低，但其对样本数量少的攻击类型和未知攻击识别率很低。我们选择 SCG 作为单个 BP 神经网络的检测结果分析，如表 2 所列，左边的 Nomal 这一栏表示 6004 条正常记录通过 SCG-BP 检测，被检测为正常的有 5538 条，被检测为 Probe 的有 60 条，被检测为 DoS 的有 403 条，被检测为 R2L 的有 3 条，所以 ANN 对正常数据的检测率为 92.2%，以此类推。从表中数据可看出，神经网络对正常数据和 DoS 的预测率较高，但是对样本数量较少以及未知攻击类型的误检率很高。

表 2 单个人工神经网络检测结果对比分析

	Nomal	Probe	DoS	U2R	R2L	其他	%
Nomal	5538	60	403	0	3	0	92.2
Probe	18	146	66	0	3	0	62.7
DoS	0	20	23461	0	0	0	99.9
U2R	0	0	3	0	0	0	0
R2L	7	6	17	0	41	0	56.9
其他	105	41	42	0	19	0	0

3.3 单个 SVM 的检测

对于 30000 条记录的测试集，如图 7 所示，SVM 有 448 个误检，误检率为 1.4933%，可见 SVM 是一种强学习器，而 BP 是一种弱学习器。类似单个 BP 神经网络，我们把 SVM 误检的记录做成表 3，可见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信息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。

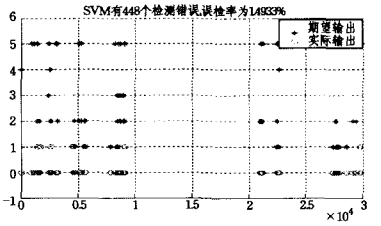

图 7 SVM 误检个数和误检率

表 3 单个 SVM 检测结果对比分析

	Nomal	Probe	DoS	U2S	R2L	其他	%
Nomal	5999	0	95	0	0	0	98.42
Probe	0	230	3	0	0	0	98.57
DoS	288	0	23193	0	0	0	98.77
U2S	1	0	0	2	0	0	64.00
R2L	0	0	2	0	70	0	97.33
其他	0	0	59	0	0	148	71.50

3.4 人工神经网络和支持向量机的集成学习检测

用集成 ANN 和 SVM 的方法对 30000 条记录的测试样本进行测试，如图 8 所示，有 54 个检测错误，误检率为 0.18%。可见，集成 ANN 和 SVM 不光误检率显著降低，而且对样本数量少的攻击和未知攻击的识别率也大大提高。表 4 统计了在入侵检测中采用各种改进的 BP 神经网络算法、SVM 以及集成 ANN 和 SVM 的检测率的比较分析。从表中看出，使用 ANN 和 SVM 的集成与目前流行的多数投票选举算法相结合的检测方法有更好的检测结果。

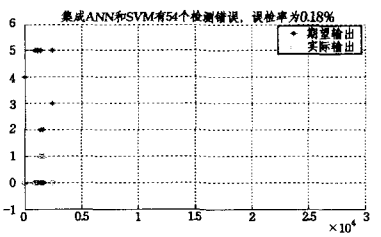

图 8 集成 ANN 和 SVM 误检个数和误检率

（下转第 224 页）

空间重构，并针对每一维时间序列采用互信息法进行延迟时间的确定，最优嵌入维数的取值方法采用最小预测误差法。根据重构相空间混沌吸引子结构，建立了多变量时间序列的局域预测模型。针对混沌局域预测法中邻近点个数少而不能满足最小二乘估计条件的问题，从减少均方误差的角度出发，引入岭估计；从消除 X 的列向量间的多重共线性关系出发，引入主成分估计，进而提出了基于正则化回归的多变量时间序列混沌局部预测模型及方法，并对实际的电力短期负荷进行了预测。实例仿真结果表明，该预测方法预测精度高，优于传统的混沌局域预测法，在邻域点数接近甚至等于嵌入维数时，预测精度仍然较佳。

参 考 文 献

[1] Stathopoulos A, Karlaftis M G. A Multivariate State Space Approach for Urban Traffic Flow Modeling and Prediction[J]. Transportation research part C, 2003, 11(2): 121-135

[2] Reick C H, Page B. Time series prediction by multivariate next neighbor methods with application to zooplankton forecasts[J]. Mathematics and Computers in Simulation, 2000, 52: 289-310

[3] Jayawardena A W, Li W K, Xu P. Neighborhood Selection for local modeling and prediction of hydrological time series[J]. Journal of Hydrology, 2002, 258(1-4): 40-57

[4] 雷绍兰. 基于电力负荷时间序列混沌特性的短期负荷预测方法研究[D]. 重庆：重庆大学，2005

[5] 方仍存，周建中，彭兵，等. 电力负荷混沌动力特性及其短期预测[J]. 电网技术，2008，32(4)：61-65

[6] 杜杰，陆金桂，曹一家. 杜短期负荷预测最大 Lyapunov 指数预报模式预测值的判定[J]. 电网技术，2006，30(20)：20-24

[7] Farmer J D, Sidorowich J J. Predicting chaotic time series[J]. Phys. Rev. Lett., 1987, 59(8): 845-848

[8] 方芬. 多变量混沌时序正则化局部线性预测[J]. 金陵科技学院学报，2007，32(2)：9-12

[9] 赵敏，樊印海，孙辉. 电力推进船舶电力负荷的多变量混沌局部预测[J]. 系统仿真学报，2008，20(11)：2797-2799

[10] 吕小青，曹彪，曾敏，等. 确定延迟时间互信息法的一种算法[J]. 计算物理，2006，23(2)：184-188

[11] 蒋传文，袁智强，侯志俭，等. 高嵌入维混沌负荷序列预测方法研究[J]. 电网技术，2004，28(3)：25-28

[12] 王海燕，盛昭瀚，张进. 多变量时间序列复杂系统的相空间重构[J]. 东南大学学报：自然科学版，2003，33(1)：115-118

[13] Cao Liangyue, Mees A, Judd K. Dynamics from Multivariate Time Series[J]. Physica D, 1998, 121: 75-88

[14] Porporato A, Ridolfi L. Multivariate Nonlinear Prediction of River Flows[J]. Journal of Hydrology, 2001, 248: 109-122

[15] Yang Hongming, Duan Xianzhong. Chaotic Characteristics of Electricity Price and its Forecasting Model[C]// IEEE CCECE 2003. Montreal, 2003: 659-662

(上接第 219 页)

表 4 各种检测算法检测结果对比分析

种类	正确率(%)				
	SCG-BP	弹性 BP	LM-BP	SVM	集成 ANN 和 SVM
正常	92.2	95.75	99.64	98.42	99.71
Probe	62.7	92.71	92.71	98.57	99.85
DoS	99.9	97.47	95.98	98.77	99.97
U2L	0.00	48.00	0.00	64.00	76.00
R2L	56.9	95.73	97.75	97.33	100.00
其他	0.00	38.65	50.24	71.50	86.47
总计	97.29	96.67	96.38	98.51	99.82

综合以上的结果，采用 ANN 和 SVM 的网络集成具有如下优点：(1)极大地提高了集成网络的泛化能力。通过集成，对各类攻击的检测率都有显著提高。(2)在各分类器均缺乏攻击全部知识的情况下，通过相互协作与知识互补，集成学习，对小样本以及未知的攻击类型都有较高的识别率。

结束语 BP 算法具有对不确定性的学习与适应能力，能很好地满足入侵检测的要求，但容易陷入局部最小，训练时间长，所以本文引入 3 种算法对 BP 算法进行改进；支持向量机在解决小样本、非线性及高维问题时具有明显优势，然而对大数据集来说，支持向量机的时空耗费非常大。综上所述，单一的学习算法往往有各自的缺陷，限制了其在入侵检测中的应用，因此本文将这些具有优良分类性能的检测器进行集成。研究表明，在集成学习中，各子分类器差异度较大时才有较好的分类正确性。我们用同一训练集对不同子分类器进行重复训练来实现集成个体的差异度，在各分类器均缺乏攻击全部知识的情况下，通过相互协作与知识互补，集成学习，来达到较高的检测率和对未知攻击类型的识别。

参 考 文 献

[1] Enning D D. An Intrusion Detection Model[J]. IEEE Trans on Software Engineering, 1987, 13(2): 222-232

[2] 杨种学，杨宁. 基于 BP 神经网络的异常入侵检测方法[J]. 计算机，2006，6：42-45

[3] 陈刚，秦茗. 基于数据挖掘的入侵检测研究[J]. 计算机仿真，2006，22(5)：43-55

[4] 王正群，陈世福，陈兆乾. 并行学习神经网络集成方法[J]. 计算机学报，2005，28(3)：402-408

[5] 饶鲜，董春曦，杨绍全. 基于支持向量机的入侵检测系统[J]. 软件学报，2003，14(4)：798-803

[6] http://kdd.ics.uci.edu/databases/kddcup99/task.htm, 2007-11-15

[7] Komorowski J O. ROSETTA—A rough set toolkit for analysis of data[C]// Fifth International Workshop on Rough Sets and Soft Computing. Tokyo, Japan, 1997: 403-407

[8] Sung A. Ranking Importance of Input Parameters of Neural Networks[J]. Expert Systems with Applications, 15: 405-41

[9] Burges C. A tutorial on support vector machines for pattern recognition[J]. Data Mining and Knowledge Discovery, 1998, 2(2): 121-167

[10] McClelland R. Parallel Distributed Processing: Explorations in the Microstructure of Cognition[M]. Cambridge, MA: MIT Press, Vol 1, 1986

[11] Schapire R E. The boosting approach to machine learning: An overview[C]// MSRI Workshop on Nonlinear Estimation and Classification. 2002

[12] Hansen L K, Salamon P. Neural network ensembles[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1990, 12: 993-1001

[13] 齐德昱，葛超，葛韧. 混合核支持向量回归及对社会用电量的预测[J]. 重庆工学院学报：自然科学版，2009，23(10)：50-52