doi:10.16652/j.issn.1004-373x.2015.23.021

基于机器学习的网络异常流量检测方法

张晓艳
（云南经济管理学院，云南 昆明 650106）

摘 要：研究一种基于机器学习的网络异常流量检测方法。使用改进型ANFIS算法作为建立的网络异常流量检测方法的核心算法。由于传统的神经网络算法使用的梯度下降算法在实际应用时，存在易陷入局部极小值，训练效率低下等问题，因此研究的改进型ANFIS算法使用附加动量算法修正模型参数，使系统能够越过误差曲面的局部最小值。最后使用KDD CUP99数据库以及LBNL实验室测试的数据对改进型ANFIS算法和BP神经网络算法的检测方法进行性能测试。结果表明，使用改进型ANFIS算法检测系统的训练效率以及检测准确率均优于使用BP神经网络算法建立的模型。

关键词：机器学习；ANFIS；BP神经网络；网络异常流量检测
中图分类号：TN711-34；TP393　　**文献标识码**：A　　**文章编号**：1004-373X(2015)23-0076-04

Research on network anomaly traffic detection method based on machine learning

ZHANG Xiaoyan
(Yunnan College of Business Management, Kunming 650106, China)

Abstract: A network anomaly traffic detection method based on machine learning is studied, in which the improved ANFIS algorithm is taken as the core algorithm. Since the gradient descent algorithm adopted by traditional neural network algorithm has the defects of easy to fall into local minimum and low training efficiency in practical application, the additional momentum algorithm is adopted by the improved ANFIS algorithm to modify the model parameters, which makes the system can cross the local minimum of the error surface. The performance of the test methods for the improved ANFIS algorithm and BP neural network algorithm is tested by using KDD CUP99 database and the data tested by LBNL laboratory. The test results show that the training efficiency and test precision of the test system with the improved ANFIS algorithm are better than that of the model established by BP neural network.

Keywords: machine learning; ANFIS; BP neural network; network anomaly traffic detection

0 引 言

随着互联网技术的飞速发展，互联网环境也逐渐变得复杂。网络中每时每刻充斥着大量的异常流量数据，威胁着人们的计算机安全，小到个人财产安全，大到国家安全。目前现实的复杂网络中的异常流量种类较多，有Alpha Anomaly异常流量、DDos异常流量、Port Scan异常流量、Network Scan异常流量等恶意行为，也有来自于如路由问题、链路故障等网络软硬件故障。因此现在对网络异常流量检测的难度也不断加大，而研究新型网络异常流量方法，应对层出不穷的网络异常流量类型，提高检测识别效率和检测准确率，已然成为现在的热点问题之一[1]。

收稿日期：2015-07-28

1 网络异常流量检测系统

1.1 网络异常流量类型

异常网络流量类型主要有Alpha Anomaly异常流量、DDos异常流量、Port Scan异常流量、Network Scan异常流量、Worms异常流量以及Flash Crowd异常流量等。可用于检测异常流量的主要流特征有目的端口总数、目的IP总数、源端口总数、源IP总数、字节数以及分组数等[2]。各个异常网络流量类型的具体含义和用于检测异常流量的流特征如下：

（1）Alpha Anomaly异常流量指高速点对点非正常数据传输行为，可用于检测Alpha Anomaly异常流量的流特征有字节数和分组数。

（2）DDos异常流量指对目标地址的分布式的拒绝服务攻击行为，可用于检测DDos异常流量的流特征有分组数、源IP地址、流计数以及目的IP地址。

(3) Port Scan异常流量指针对容易受到网络攻击的主机端口的扫描,可用于检测Port Scan异常流量的流特征有分组数、源端口以及源IP地址。

(4) Network Scan异常流量指针对不同的网络地址的同一个端口的扫描行为,可用于检测Network Scan异常流量的流特征有分组数、源IP地址、流计数、目的端口以及目的IP地址。

(5) Worms异常流量实际属于一种特殊的Network Scan异常流量,指在网络中利用网络安全的漏洞而进行自身复制,可用于检测Worms异常流量的流特征有目的端口以及目的IP地址。

(6) Flash Crowd异常流量指对于某一个资源或者服务的大量非正常用户的请求,可用于检测Flash Crowd异常流量的流特征有源IP地址、目的端口、分组数、目的IP地址以及流计数[3-7]。

1.2　基于机器学习的网络异常流量检测方法

机器学习智能算法能够有效解决各种识别问题,在识别检测领域已经得到了广泛应用。机器学习智能算法一般分为有督导机器学习和无督导机器学习[8-9]。

基于有督导机器学习算法的分类识别模型的一般核心思想是使用已知确定类型的样本数据对识别模型进行机器学习并建立对应的分类规则,并根据分类规则对未知的未确定类型的样本数据进行分类识别。基于有督导机器学习算法的分类识别模型的优点是检测识别率较高,缺点是需要大量已知类型数据样本对模型进行训练以使得模型具有较好的泛化能力,并且无法对未确定类型的样本数据进行分类识别。有督导机器学习算法一般有贝叶斯算法、决策树算法、支持向量机算法以及神经网络算法等[10]。

基于无督导机器学习算法的分类识别模型的一般核心思想是使用样本数据的特征相似度来聚合分簇,以得到各簇和类的映射。基于无督导机器学习算法的分类识别模型的优点是能够对未确定类型的样本数据进行分类识别,但是缺点同样明显,就是识别分类的速度和准确度比较低[11]。

本文着重研究基于神经网络机器学习算法的网络异常流量检测方法。

2　ANFIS算法模型

ANFIS算法模型通常有5个网络层,分别是1个输入层和输出层、2个规则层,结构如图1所示。

在第一个网络层中,对输入变量进行模糊化,各个节点的输出为:

$$O_i^1(x_1)=\mu_{A_i}(x_1),\ \ i=1,2 \tag{1}$$

$$O_j^1(x_2)=\mu_{B_j}(x_2),\ \ j=1,2 \tag{2}$$

式中:x_1, x_2是输入的节点;O_i^1, O_j^1是输出的节点。

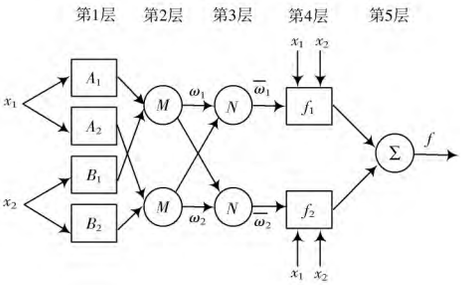

图1　ANFIS算法结构

在第二个网络层中,将输入变量相乘,各个节点输出值表示规则强度为:

$$O_i^2=\omega_i=\mu_{A_i}(x_1)\mu_{B_i}(x_2),\ \ i=1,2 \tag{3}$$

在第三个网络层中,归一化处理规则的强度为:

$$O_i^3=\overline{\omega}_i=\frac{\omega_i}{\omega_1+\omega_2},\ \ i=1,2 \tag{4}$$

在第四个网络层中,计算各个模糊规则输出值:

$$O_i^4=\overline{\omega}_i f_i=\overline{\omega}_i(p_i x_1+q_i x_2+r_i),\ \ i=1,2 \tag{5}$$

在第五个网络层中,计算节点的输出值:

$$O_i^5=f=\sum_{i=1}^{2}\overline{\omega}_i f_i=\frac{\sum_{i=1}^{2}\omega_i f_i}{\sum_{i=1}^{2}\omega_i} \tag{6}$$

ANFIS算法的总输出通常由给定的前提、结论参数得到:

$$\begin{aligned} f&=\frac{\omega_1}{\omega_1+\omega_2}f_1+\frac{\omega_2}{\omega_1+\omega_2}f_2\\ &=\overline{\omega}_1 f_1+\overline{\omega}_2 f_2\\ &=\overline{\omega}_1 x_1 p_1+\overline{\omega}_1 x_2 q_1+\overline{\omega}_1 r_1+\overline{\omega}_2 x_1 p_2+\overline{\omega}_2 x_2 q_2+\overline{\omega}_2 r_2 \end{aligned} \tag{7}$$

ANFIS算法模型主要是使用混合算法对前提和结论参数不断更新。通常将一个初始值赋予给前提参数,结论参数由最小二乘估计算法得到。最终从最后一层反向向第一层由梯度下降算法传递系统的误差,以不断更新前提参数[12]。

由于梯度下降算法在实际应用时,存在易陷入局部极小值,训练效率低下等问题,因此本文研究的改进型ANFIS算法使用附加动量算法来修正模型参数,使系统能够越过误差曲面的局部最小值。附加动量算法的具体形式为:

$$c_i(n+1)=c_i(n)+\Delta c_i(n) \tag{8}$$

$$\Delta c_i(n)=\lambda\Delta c_i(n-1)-(1-\lambda)\beta(n)\frac{\partial E(n)}{\partial c_i(n)} \tag{9}$$

$$\sigma_i(n+1)=\sigma_i(n)+\Delta\sigma_i(n) \tag{10}$$

$$\Delta\sigma_i(n)=\lambda\Delta\sigma_i(n-1)-(1-\lambda)\beta(n)\frac{\partial E(n)}{\partial\sigma_i(n)} \tag{11}$$

式中:λ是动量因子,一般取0.95左右;n是迭代步数;

$\beta(n)$ 是第 n 步运算的学习率[13]。

3 实验分析

3.1 实验数据采集

使用VB.net配合MySQL数据库软件建立网络异常流量检测系统。用于网络异常流量检测模型训练和测试的数据来源于Mitlincoln 实验室的KDD CUP99数据库以及LBNL实验室测试的数据。从数据库中随机抽取Alpha Anomaly异常流量、DDos异常流量、Port Scan异常流量、Network Scan异常流量、Worms异常流量以及Flash Crowd异常流量各200条数据样本，其中100条数据样本作为训练样本，另外100条数据样本作为测试样本。

使用上述数据库得到的网络异常流量数据包含了异常流量的主要流特征，如目的端口总数、目的IP总数、源端口总数、源IP总数、字节数以及分组数等。由于各个流特征取值以及使用的度量单位均不相同，因此必须对数据进行归一化处理：

$$x'_i=\frac{x_i-\bar{x}}{S},\ i=1,2,\cdots,n \tag{12}$$

式中：x' 是归一化处理后的数据，数值在0~1之间；$\bar{x}$ 是流特征数据的数值平均值，$\bar{x}=\left(\sum_{i=1}^{n}x_i\right)\Big/n$；$S$ 是样本的特征标准差，$S=\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2}$；$n$ 是流特征数据样本的数量[14]。

3.2 检测系统模型建立

为了简化检测系统的模型，使用减法聚类算法对归一化处理后的流特征数据样本空间进行非线性规划，使用生成的Sugeno型结构作为网络异常流量检测模型的初始结构，对检测模型的各个参数使用混合学习算法以及附加动量算法逐步优化。所建立的ANFIS模型选用三角函数型的隶属度函数，ANFIS模型的 a,b,c 参数学习率设定为0.01，误差上限为 10^{-3}。可以得到训练误差曲线如图2所示。

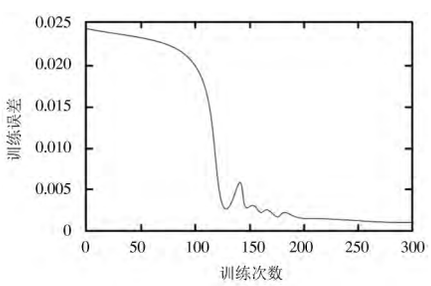

图2 基于ANFIS的模型训练误差

为了对比本文研究的ANFIS模型的优势，使用BP神经网络进行对比实验。BP神经网络模型的输入层单元数根据流特征数设定为6，输出层单元数根据网络异常流量种类设定为6，隐含层单元数根据经验公式设定为11，使用Levenberg-Marquardt算法。BP神经网络模型的训练误差曲线如图3所示。

通过对比模型训练误差可以看出，ANFIS算法比BP神经网络算法使用的训练时间更短。使用附加动量算法修正模型参数，使系统能够越过误差曲面的局部最小值[15]。

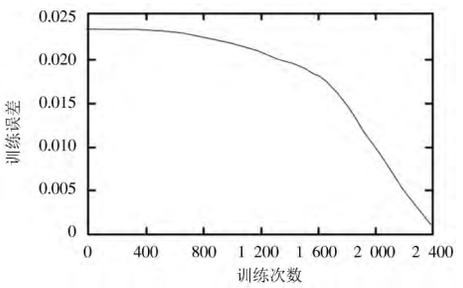

图3 基于BP神经网络的模型训练误差

3.3 检测系统测试分析

网络异常流量监测过程如图4所示。

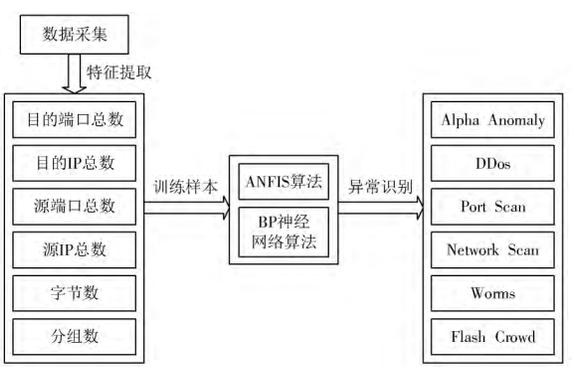

图4 模型训练误差

为了评判本文建立的各个网络异常流量检测模型的性能，使用准确率和误报率进行评价[16]。

$$\text{准确率}=\frac{\text{正常检测出异常流量的个数}}{\text{异常流量的总数}}\times 100\% \tag{13}$$

$$\text{误报率}=\frac{\text{把正常流量误判为异常流量的个数}}{\text{样本总数}}\times 100\% \tag{14}$$

使用本文提出的ANFIS算法和使用BP神经网络算法建立的网络异常流量检测模型的测试结果如图5和图6所示。其中纵、横轴坐标表示的含义如表1所示。

使用本文提出的ANFIS算法和使用BP神经网络算法建立的网络异常流量检测模型，对测试数据进行检测得到的准确率和误报率对比如表2所示。

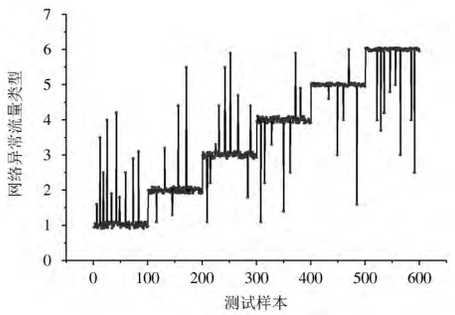

图5　基于ANFIS算法模型的测试结果

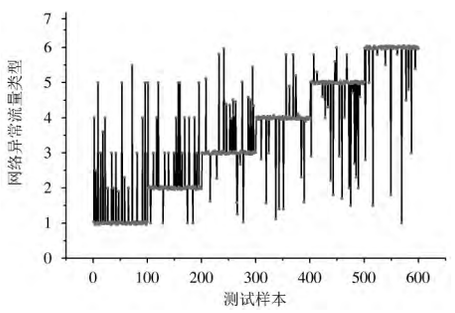

图6　基于BP神经网络算法模型的测试结果

表1　图5和图6的横、纵坐标含义

横轴	含义	纵轴	含义
1~100	Alpha Anomaly异常流量的测试样本	1	判断为Alpha Anomaly异常流量的样本
101~200	DDos异常流量的测试样本	2	判断为DDos异常流量的样本
201~300	Port Scan异常流量的测试样本	3	判断为Port Scan异常流量的样本
301~400	Network Scan异常流量的测试样本	4	判断为Network异常流量的样本
401~500	Worms异常流量的测试样本	5	判断为Worms异常流量的样本
501~600	Flash Crowd异常流量的测试样本	6	判断为Flash Crowd异常流量的样本

表2　两种方法的准确率和误报率结果　%

异常流量类别	改进ANFIS		BP神经网络	
	准确率	误报率	准确率	误报率
Alpha Anomaly	90.2	9.8	83.7	16.3
DDos	95.8	4.2	79.2	20.8
Port Scan	89.7	10.3	80.6	19.4
Network Scan	94.4	5.6	75.7	24.3
Worms	96.2	3.8	79.3	20.7
Flash Crowd	91.9	8.1	78.2	21.8

通过对比两种检测方法得到的准确率和误报率可以看出,本文提出的改进型ANFIS算法相比使用BP神经网络算法使得网络异常流量检测系统具有更高的准确率和误报率,能够有效避免使用BP神经网络容易陷入局部最小值以及收敛速度低等问题。

4　结　语

本文研究了一种基于机器学习的网络异常流量检测方法。首先对网络异常流量分类及识别检测方法进行了研究。由于传统的BP神经网络存在容易陷入极小值并且训练速度慢等问题,本文提出使用ANFIS算法建立异常流量检测识别模型。最后通过实验来验证使用ANFIS算法建立模型的识别性能,其性能要优于传统BP神经网络算法的异常流量检测识别模型。

参　考　文　献

[1] 李天枫,王劲松,王立学.基于IPFIX的大规模网络异常流量检测机制研究[J].天津理工大学学报,2015(3):1-5.

[2] 王涛,余顺争.基于机器学习的网络流量分类研究进展[J].小型微型计算机系统,2012(5):1034-1040.

[3] 刘俊利.基于ANFIS的多信息融合煤岩识别方法研究[J].中国煤炭,2014(12):56-59.

[4] 孙丙香,高科,姜久春,等.基于ANFIS和减法聚类的动力电池放电峰值功率预测[J].电工技术学报,2015(4):272-280.

[5] 姚宏林,韩伟杰,吴忠望.基于模糊相对熵的网络异常流量检测方法研究[J].信息安全与技术,2014(8):16-18.

[6] 赵鑫.基于NetFlow的网络流量异常检测技术研究[D].保定:河北大学,2014.

[7] 李洪洋.浅析网络异常流量分析检测研究与实现[J].网络安全技术与应用,2013(10):63-64.

[8] 穆祥昆,王劲松,薛羽丰,等.基于活跃熵的网络异常流量检测方法[J].通信学报,2013(z2):51-57.

[9] 燕发文,黄敏,王中飞.基于BF算法的网络异常流量行为检测[J].计算机工程,2013(7):165-168.

[10] 吴小花.网络异常流量识别技术的研究[D].长春:长春工业大学,2013.

[11] 申磊.基于机器学习的异常流量检测系统研究[D].北京:北京邮电大学,2013.

[12] 周丹,南敬昌,高明明.改进的简化粒子群算法优化模糊神经网络建模[J].计算机应用研究,2015,32(4):1000-1003.

[13] 秦也辰,管继富,顾亮,等.基于自适应神经模糊网络的路面识别技术[J].北京理工大学学报,2015(5):481-484.

[14] 姚晔.量子粒子群和最小二乘支持向量机相结合的网络异常检测[J].微电子学与计算机,2012(3):39-42.

[15] 许倩,程东年,程国振.一种半监督联合模型下的异常流量检测算法[J].小型微型计算机系统,2013(6):1242-1247.

[16] 李春林,黄月江,牛长喜.一种面向云计算的网络异常流量分组方法[J].计算机应用研究,2014(12):3704-3706.

作者简介:张晓艳(1983—),女,云南永胜人,硕士,讲师。主要从事现代教育技术、计算机科学、软件工程等研究。