DOI:10.19329/j.cnki.1673-2928.2019.04.012

基于深度学习的实时网络入侵检测方法

朱平哲
(三门峡职业技术学院 信息传媒学院,河南 三门峡 472000)

摘 要:网络安全问题日益严重,能够快速、有效地发现各类入侵行为的入侵检测技术成为当前的研究热点。本文汲取机器学习的性能优势,提出了一种基于深度学习的实时网络入侵检测方法,通过对带有入侵数据的大量训练样本的学习,构建用于区分正常状态和入侵状态的入侵检测模型。为了测试和评估该方法的性能,将其与随机森林、Logistic回归和贝叶斯等其他机器学习模型进行比较。实验结果表明,基于H2O深度学习多项式模型的性能均优于其他机器学习模型,可良好应用于实时网络入侵检测。

关键词:入侵检测;实时网络;深度学习;多项式模型

中图分类号:TP311.1 **文献标志码**:A **文章编号**:1673-2928(2019)04-0048-04

0 引言

随着计算机和网络技术的飞速发展,网络安全问题日益受得人们的关注,它已然成为一个全球性的重大信息安全问题。由于信息共享的程度不断提高,人们在使用网络提供的各种服务和信息的同时,信息的安全也受到各种恶意行为和攻击的严重威胁,面临着日益增加的网络入侵的困扰,网络安全问题成为迫切需要解决的问题之一。据美国《金融时报》报道,现在平均每20秒就发生一次入侵计算机网络的事件,超过1/3的互联网防火墙被攻破。计算机网络已经成为关键的基础设施,涉及政府机构、军事部门、科研院校、金融商业等部门的计算机犯罪,严重干扰了人们的日常生活,造成巨大经济损失,甚至直接或间接地威胁到国家安全。

入侵检测是一种具有主动防御能力的安全防护技术,它通过相关技术及时地检测出可能发生的入侵行为,从而大大提高目标系统的安全防护能力[1]。但现有的入侵检测系统大多存在检测时间较长、检测精度低、误报率和漏报率高等不足。基于机器学习的入侵检测是网络安全领域研究的热点,它通过对带有入侵数据的大量训练样本的学习,构建一个用于区分正常状态和入侵状态的入侵检测模型。但目前仍然存在着许多有待解决的问题,如建立分类器模型所需要的训练样本过多、训练样本标注耗费大量时间且过分依赖于专业知识等问题。

深度学习作为一种新的解决方案已经出现,由于其具有诸如前馈和反向传播等优势,因而它可提供更有效的网络入侵检测的潜力[2]。

鉴于此,本文深入研究了如何利用深度学习模型原理建立实时网络入侵检测系统,提出一种基于深度学习的实时网络入侵检测方法。该方法结合了深度学习二项分类模型来预测是否存在入侵,进而使用深度学习多项式模型来识别入侵类别。为了提供消息传递服务该深度学习模型在建立时选择的是H2O深度学习库构建其模型。最后为测试该模型的有效性及性能评估使用NSL-KDD数据集[3]进行了评价研究,并将H2O深度学习的二项式/多项式模型和随机森林、Logistic回归、贝叶斯等其他机器学习模型进行比较和分析。

1 基于深度学习的入侵检测

近年许多学者对基于深度学习的入侵检测方法进行了研究和探索。Kim[3]等利用深层神经网络以及对基本数据的清理和复制消除了特征对数据和训练数据进行转换的预处理过程。由于KDD-cup99数据集包含整数和浮点数,因此将所有数据转换为字符串类型防止最小化数据丢失,进而利用10%的校正数据训练模型,从而达到了准确率99%和误报率0.08%的良好效果。Saxe和Berlin[5]采用了大于400k的二进制文件来构建三层深度学习模型。第一层是训练层,第二层是对于标记数据和未标记数据的数据分类层,而最后一层是用于全部数据分类。经反复实验,这种模型获得了

收稿日期:2018-12-26
基金项目:河南省教育厅科学技术研究重点项目(15B520026)。
作者简介:朱平哲(1982-),女,河南驻马店人,讲师,硕士,研究方向:智能信息处理。

(C)1994-2019 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

95%的准确率，但其缺点为缺乏数据清理这一重要过程，因为KDDcup99数据集包含超过70%的重复记录，因而数据清理的过程必不可少[4]。

Lecun等[5]采用了一个新型的神经网络，它是以顺序模式获取输入。当每行通过系统时，它保持前一个结果的向量状态，并用前一个模型进行训练。由于向量梯度值在训练阶段可能压缩或中断，所以训练过程很困难。李阳等[6]使用自动编码器方法和相同的基准数据集KDDcup99来训练和测试模型。自动编码器、深度置信网络和归一化训练数据可同时使用，以减少日志丢失。基于自动编码器的方法和归一化方法均给出了比递归神经网络方法更好的结果。虽然这两种方法都提供了良好的结果，但它们不具有实时性。

Salama等[2]实现了受限玻尔兹曼机系统。它是在结合已有的SVM算法和深度置信网络的基础上进行工作的。输入数据集流经提取13个网络参数的系统，这些参数被SVM用作输入，以将输入分类为入侵或不入侵。Niyaz等[8]提出了自学习的三个阶段。特征学习是从使用未标记数据的第一层开始。第二阶段是将学习向量应用于标记数据集，然后进行软最大回归和分类从而预测网络入侵检测。软最大回归模型与其他回归模型相比，具有更高的效率且易于实现。在Matlab系统上采用三级数据清理流程以减少日志丢失。在测试数据和训练中，他们分别获得了98.84%和88.39%的F度量值。因此他们得到的下一步重要研究目标就为将该模型转化为实时入侵检测系统。

综上所述，以上针对入侵检测问题的解决方法均是基于深度学习模型的。这些方法分别给出了高低不同的性能水平，也都存在一些缺点和其他的困难。因此，本文研究主要是在深度学习性能提高和实时性应用两个方面为解决网络入侵检测问题。由此本文提出一种基于深度学习的实时网络入侵检测方法。

2 基于深度学习的实时入侵检测模型

深度学习在实时入侵检测中的性能研究及方法模型仍在初步探索阶段，因而本文将以实时入侵检测为研究重点，尝试将深度学习二项式分类模型与深度学习多项式模型相结合的方法检测和预测入侵，以识别攻击类别。

2.1 数据集

大量研究文献及报告显示，大多数与入侵检测相关的研究与实验都使用了基准KDDcup99数据集。然而，使用该数据集的缺点在于有可能会影响模型训练的重复值。KDcup99训练和测试数据集中分别包含有78%和75%的重复记录[3]，因此即使是基本的机器学习模型也可在训练数据上达到98%以上的准确率，在测试数据上达到86%以上的准确率。

因此，本文的研究则使用在KDDcup99数据集中由选定记录组成的NSL-KDD数据集，因为该数据集不具有重复值高的缺点[3]。NSL-KDD训练数据集具有41个属性、3个标识型属性（即协议、服务、标志）和38个数值型属性（如持续时间、源字节、目的地字节、错误片段、失败登录次数等）。该数据集是由正常流量和39种攻击类型组成，这些攻击类型被分为4组，即拒绝服务（DoS）、探测（Probe）、远程到本地（R2L）和用户到远程（U2R）[7]。

NSL-KDD数据集为一种结构化格式，是KDDcup99数据集的一个简明示例，因此只需要很少的预处理。检测训练数据和测试数据集是否有缺失值，并将完成所有记录，因此不需要消除记录或进行归类。一些机器学习算法只处理数值，因此标识型属性必须采用编码映射为多个二进制的数值属性。然而，通过H2O深度学习模型可实现自动完成其映射。

2.2 H2O深度学习模型

本研究受Niyaz[7]、Arora[8]等研究成果的启发，深度学习所具备的学习新特征使其具有高精度和较高的准确率，由此在机器学习的诸多模型中基于深度学习的神经网络模型的性能是最佳的。因此，本文提出了网络入侵检测的深度学习方法，并进一步将功能扩展至实时入侵检测模型。

本文研究的网络入侵检测系统是分别基于两种深度学习模型建立的。第一种模型是基于深度学习的二项式分类模型，适用于正常的网络入侵预测。第二种模型是基于深度学习的多项式模型，用于第一模型检测到入侵后进一步检测其入侵行为类别的（包括DoS、Probe、R2L和U2R等入侵行为）。

在深度学习的诸多模型中，本文选择使用H2O深度学习，即使用开源H2O库开发深度学习模型。选择H2O深度学习模型的主要原因是其对机器学习的广泛接受以及它向Web应用程序提供基于POJO的API服务。由于Web应用模型在网络管理方面是至关重要的，因此基于Web的应用模型是系统的重要组成部分。在创建H2O深度学习模型后，即可下载Java POJO类。下载模型的方法有两种，使用本地主机Web UI或H2O库下载。此外，H2O深度学习模型还提供了包含所有依赖性JAR的H2O生成模型以支持Java API。在这种情况下，一个用于二项分类模型或多项式模型的Java类可与H2O生成模型JAR同时下载。

(C)1994-2019 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

2.3 H2O深层学习模型的入侵检测实验

本文以我校园网的实际网络流量为实验数据，基于深度学习二项式分类模型的网络入侵检测系统检测的网络预测的实验结果数据集如表1所示，基于深度学习多项式模型的网络入侵检测系统进一步检测其入侵行为类别的实验结果如表2所示。

表1 正常网络流量和不同网络攻击类别的分布数据

类别	训练数据		测试数据	
	网络流量数量	正常和异常比率/%	网络流量数量	正常和异常比率/%
正常流量	67 343	53.46	9 710	43.07
异常流量	58 630	46.54	12 833	56.93
DoS	45 927	36.46	7 460	33.09
Probe	11 656	9.25	2 421	10.74
R2L	995	0.79	2 885	12.80
U2R	52	0.04	67	0.30
总数量	125 973	100.00	22 543	100.00

表2 正常网络流量和不同网络攻击类别的分布数据

训练数据		测试数据	
网络流量数量	正常和异常比率/%	网络流量数量	正常和异常比率/%
正常流量	53.46	正常流量	43.07
neptune	32.72	neptune	20.66
satan	2.88	guess passwd	5.46
ipsweep	2.86	mscan	4.42
portsweep	2.33	warezmaster	4.19
smurf	2.10	apache2	3.27
nmap	1.19	satan	3.26
back	0.76	processtable	3.04
teardrop	0.71	smurf	2.95
warezclient	0.71	back	159
pod	0.16	snmpguess	1.47
其他异常行为	0.12	其他异常行为	6.62

2.4 两种模型的性能评估实验

为了测试两种模型的性能，本文将与随机森林、Logistic回归和贝叶斯等其他经典的机器学习模型进行比较实验。由于所有这些机器学习库都在Java虚拟机中工作，因而可更为方便直接地进行性能比较。

本实验采用两种方法对模型进行性能评估。第一种方法是对NSL-KDD训练数据进行5倍交叉验证。第二种方法在没有验证分割的情况下对训练数据集上的模型进行训练，并在测试数据集上对模型进行测试。评估机器学习模型的性能度量指标主要有5个，包括准确度、精度、F-测量值、AUC(用于评估鲁棒性)、检测率(正确分类为属于特定类实例的比率)。

针对本文所提出的两种模型、随机森林、Logistic回归和贝叶斯等5种机器学习模型进行NSL-KDD训练数据集的5倍交叉验证法实验，得到的性能评估度量结果如表3所示。针对本文所提出的两种模型、随机森林、Logistic回归和贝叶斯等5种机器学习模型进行NSL-KDD测试数据集的无交叉验证法实验，得到的性能评估度量结果如表4所示。两次实验结果表明，本文所提的两种深度学习模型中，基于深度学习的多项式模型性能更好，且与其他三种典型机器学习模型相比，综合性能也有一定的优势。

3 结束语

本文提出了一种基于深度学习的实时网络入侵检测方法，该方法给出了两种机器学习模型——H2O深度学习的二项式分类模型和H2O深度学习的多项式分类模型。实验结果表明，与其他三种机器学习模型相比，基于H2O深度学习的多项式分类模型的性能良好，各性能度量指标值均有一定的优势。因此，本文所提的方法为网络入侵检测问题提供了一种新的尝试。

表3 5种机器学习模型的NSL-KDD训练数据集的5倍交叉验证实验

机器学习模型	准确度/%	精度/%	F-测量值	AUC	正常流量检测率/%	入侵行为检测率/%
多项式深度学习	99.52	99.95	99.50	99.94	99.66	99.43
二项式深度学习	96.79	96.80	96.80	99.00	97.05	96.48
随机森林	98.91	97.99	99.90	100.00	99.66	98.24
Logistic回归	97.08	97.10	97.10	99.40	98.22	95.77
Naïve-Bayes	90.34	90.40	90.30	96.60	93.59	86.65

(C)1994-2019 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

表4 5种机器学习模型的NSL-KDD训练数据集的无交叉验证实验

机器学习模型	准确度/%	精度/%	F-测量值	AUC	正常流量检测率/%	入侵行为检测率/%
多项式深度学习	83.87	81.47	81.83	87.42	97.27	84.20
二项式深度学习	76.12	80.20	76.00	75.10	91.30	64.64
随机森林	80.25	81.60	80.40	88.00	92.55	81.37
Logistic回归	74.58	85.10	74.30	75.50	90.15	80.22
Naïve-Bayes	76.11	79.80	75.90	85.66	87.66	78.68

参考文献:

[1] WANG L, JONES R. Big Data Analytics for Network Intrusion Detection:A Survey[J]. International Journal of Networks and Communications, 2017, 7(1): 24-31.

[2] SALAMA M A, EID H F, RAMADAN R A, et al. Hybrid Intelligent Intrusion Detection Scheme[M]. Soft Computing in Industrial Applications, Springer, 2011.

[3] TAVALLAEE M, BAGHERI E, LU W, et al. A detailed analysis of the KDD CUP 99 data set[C]. 2009 IEEE Symposium on Computational Intelligence for Security and Defense Applications (CISDA), Jul. 2009(1): 1-6.

[4] KIM J, SHIN N, JO S Y, et al. Method of intrusion detection using deep neural network[C]. 2017 IEEE International Conference on Big Data and Smart Computing (BigComp), Feb., 2017(1): 313-316.

[5] Y. Le Cun, Y. Bengio, and G. Hinton. Deep learning[J]. Nature, 2015, 521(7553): 436-444.

[6] 李阳.一种基于深度学习的混合恶意代码检测方法[J].安全及其应用,2015,9(5):205-216.

[7] 王奇安,陈兵.基于广泛内核的CVM算法的入侵检测[J].计算机研究与发展,2017,49(5):974-982.

[8] 努尔布力,柴胜,李红炜,等.一种基于Choquet模糊积分的入侵检测警报关联方法[J].电子学报,2017,39(12):2741-2747.

A Real-time Network Intrusion Detection Technique Based on Deep Learning

ZHU Pingzhe

(Informational Media Department, Sanmenxia Polytechnic, Sanmenxia City, Sanmenxia 472000,China)

Abstract:The problem of network security is becoming more and more serious. Intrusion detection technology, which can quickly and effectively detect all kinds of intrusions, has become a research hotspot. Drawing on the performance advantages of machine learning, this paper proposes a real-time network intrusion detection technique based on deep learning. By learning a large number of training samples with intrusion data, an intrusion detection model is constructed to distinguish normal state from intrusion state. In order to test and evaluate the performance of this method, this paper compares it with other machine learning models such as random forest, logistic regression and Bayesian. The experimental results show that the performance of the polynomial model based on H2O deep learning is better than that of other machine learning models, and it can be well applied to real-time network intrusion detection.

Key words:intrusion detection; real-time network; deep learning; polynomial model

(责任编辑:郝安林)

(C)1994-2019 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net