计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.200700129

# 基于 Spark 的车联网分布式组合深度学习入侵检测方法

俞建业[1]　戚　涌[1]　王宝茁[2]
1 南京理工大学计算机科学与工程学院　南京 210094
2 江苏中达智能交通产业研究院有限公司　江苏 常州 213000
(365633995@qq.com)

**摘　要**　随着 5G 等技术在车联网领域中被广泛应用,入侵检测作为车联网信息安全重要的检测工具发挥着越来越重要的作用。由于车联网结构变化快,数据流量大,入侵形式复杂多样,传统检测方法无法确保其准确性和实时性要求,不能直接被应用到车联网。针对这些问题,提出了一种基于 Apache Spark 框架的车联网分布式组合深度学习入侵检测方法,通过构建 Spark 集群,将深度学习卷积神经网络(Convolutional Neural Networks,CNN)和长短期记忆网络(LSTM)组合,进行车联网入侵特征提取和数据检测,从大规模车联网数据流量中发现异常行为。实验结果证明,与其他现有模型相比,该模型算法在时间上最快达到 20.1s,准确率最高可达 99.7%,具有较好的检测效果。

**关键词**:入侵检测;车联网;CNN;LSTM;Apache Spark
**中图法分类号**　TP302.7

## Distributed Combination Deep Learning Intrusion Detection Method for Internet of Vehicles Based on Spark

YU Jian-ye[1],QI Yong[1] and WANG Bao-zhuo[2]
1 School of Computer Science and Engineering,Nanjing University of Science & Technology,Nanjing 210094,China
2 Jiangsu Zhongda Intelligent Transportation Industry Research Institute Co. Ltd,Changzhou,Jiangsu 213000,China

**Abstract**　With the application of 5G and other technologies in the field of Internet of vehicles,intrusion detection as an important detection tool for information security of Internet of vehicles plays an increasingly important role. Due to the rapid change of the structure of the Internet of vehicles,large data flow,complex and diverse forms of intrusion,traditional detection unable ensure the accuracy and real-time requirements,and unable be directly applied to the Internet of vehicles. To solve these problems,this paper proposes a distributed combination deep learning intrusion detection method for Internet of vehicles based on Apache spark framework. By constructing spark cluster,the deep learning CNN and LSTM are combined to extract intrusion features and detect data,and find abnormal behaviors from large-scale Internet of vehicles data traffic. Experimental results show that,compared with other existing models,the proposed method can achieve 20.1s in time and 99.7% in accuracy.

**Keywords**　Intrusion detection,Internet of vehicles,CNN,LSTM,Apache Spark

## 1　引言

随着新兴技术在车联网领域的实践应用,车联网得到了更加快速的发展,但因其自身的特殊性,即汽车本身对网络安全的考虑不足、车载计算机的能力受限、复杂的应用环境、大量的分布式节点以及传感网络,这些都需要极高的安全要求。车联网的安全问题越来越成为其落地应用的绊脚石,如何确保车联网安全体系中车-路-云通信安全以及识别各种恶意攻击行为已成为业内人士和信息安全专家密切关注的重点。入侵检测(Intrusion Detection,ID)是一种用于检测任何通信系统中的入侵者和攻击的网络安全技术,通过识别或检测各种攻击行为,监视和分析网络流量以对正常和异常行为进行分类,识别网络中出现的威胁等异常活动。该技术作为一种主动防御技术,成为了保障车联网安全关键机制之一。

将机器学习算法应用于传统互联网入侵检测系统是目前主流的研究方向。Halima 等[1]将 SVM 方法应用到入侵检测系统(Intrusion Detection System,IDS)中。采用 SVM 和 Naive Bayes 机器学习算法,应用归一化和特征简约进行分析对比。但是基于机器学习的入侵检测机制的重要缺点是需要大量的训练时间来处理网络先前数据流的大型数据集,在处理大数据网络环境中,深度学习技术具有良好的自我学习功能、联想存储功能以及高速寻优功能,非常适用于处理目前复杂

基金项目:国家重点研发计划政府间国际科技创新合作重点专项(2016YFE0108000);工业和信息化部网络安全技术应用试点示范项目:智能网联车路协同通信安全研究应用平台;江苏省重点研发计划(产业前瞻与共性关键技术)项目(BE2017163);江苏省交通运输科技项目(2018Y45)
This work was supported by the National Key Research and Development Program Intergovernmental Key Items for International Science and Technology Innovation Cooperation of China(2016YFE0108000),Ministry of Industry and Information Technology Network Security Technology Application Pilot Demonstration Project:Intelligent Networked Vehicle-road Cooperative Communication Safety Research Application Platform,Key Research and Development Program of Jiangsu Province China(BE2017163) and JiangSu Transportation Technology Project(2018Y45).
通信作者:戚涌(790815561@qq.com)

的网络流量数据，尤其是复杂的车联网环境。

目前，已有较多基于深度学习和大数据分布式技术的入侵检测的研究。Vinayakumar 等[2]提出混合深度神经网络(DNN)模型用于检测和分类未知的网络攻击。Dong 等[3]认为深度学习近年来得到广泛关注，他将传统的经典方法与新型深度学习方法进行了比较。Ishaque 等[4]使用深度学习的智能功能来构建智能入侵检测系统。Yao 等[5]提出了一种基于混合 MLP/CNN 神经网络的异常入侵检测方法。Wang 等[6]提出了基于深度学习的网络入侵检测方法，通过使用 BP 神经网络对入侵类型进行分类，对 KDD-CUP99 数据集进行了验证。Ding 等[7]提出基于深度卷积神经网络的入侵检测方法，将网络数据转换为图像并进行降维。通过训练和识别来提高检测的准确率、误报率和检测速率。Chockwanich 等[8]通过 TensorFlow 顶部使用 Keras 将受监督的深度学习对不同的攻击进行分类，得到 RNN 深度学习技术具有最高的准确性。Dobson 等[9]利用 Spark 框架来实现随机森林、SVM 等机器学习算法，并与深度多层感知器进行比较。通过研究可以发现，深度学习算法在准确率上比传统的机器学习算法高，但在分析数据时所花的时间更长。

传统静态网络入侵检测一般分为基于主机的入侵检测和基于网络的入侵检测。而车联网入侵检测是通过过滤车辆之间的交换数据来实现的。由于车联网也是连接到互联网或其专用的网络中，其传统的恶意攻击手段对车联网同样有效，而且破坏性更大，这对入侵检测的安全性提出了更高的要求。结合车联网网络大流量和多维复杂性的特点，利用深度神经网络检测方面的优势和分布式并行计算的快速有效特性，本文提出将组合深度学习算法应用于 Spark 框架[10-11]中进行入侵检测。通过搭建基于 Spark 架构集群，对传统的深度学习算法进行改进，将 CNN 与 LSTM[12]联合运用，提出 CNN-LSTM 算法模型，用于分析 NSL-KDD 数据集[13]和 UNSW-NB15 数据集[14]，以最大限度地减少对连接车辆的安全攻击，其主要目标是使检测攻击所需的时间最小化，并提高分类任务的准确性，更加适用于车联网的实际环境。通过实验分析发现，各个指标均有较好的提升。

## 2 分布式组合深度学习入侵检测方法

### 2.1 总体架构

本文提出基于 ApacheSpark 框架下的车联网分布式组合深度学习入侵检测方法，其总体架构如图 1 所示。

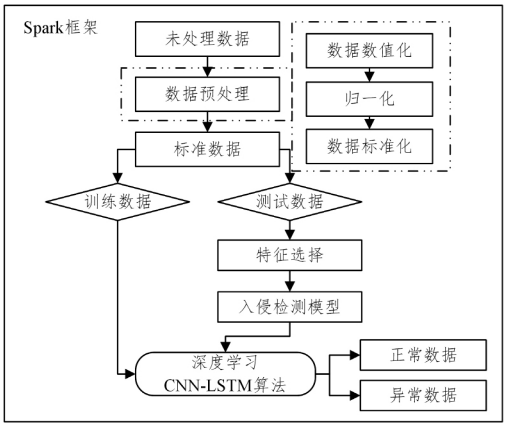


图 1 Spark 框架入侵检测方法的总体架构

Fig. 1 General architecture of Spark framework intrusion detection method

Spark 框架入侵检测方法包含 3 个步骤：

Step1 数据预处理。将数据集带符号属性的参数进行数值化和归一化处理，得到标准数据集。

Step2 数据训练。对标准化数据集进行训练和调值处理。

Step3 分类处理。基于深度学习 CNN-LSTM 算法，对测试的数据集进行测试识别，得到结果。

### 2.2 CNN-LSTM 组合算法

CNN 适合提取数据特征，LSTM 适用于处理时间序列，解决时间序列数据之间的依赖性问题，提高识别的准确性。本文结合两种算法的优点，提出了 CNN-LSTM 算法。

卷积神经网络(CNN)[15]由多层感知机(MLP)[16]演变而来，与传统的特征选择算法相比，该算法能更好地自动学习特征，而且流量数据越多，CNN 能够学习到的有用特征就越多，分类就越好，非常适合大规模的网络环境。如图 2 所示，其结构分为卷积层、池化层以及全连接层。卷积层的作用是提取特征；池化层的作用是对特征进行抽样。最后全连接层负责把提取的特征连接起来，通过分类器得到分类结果。

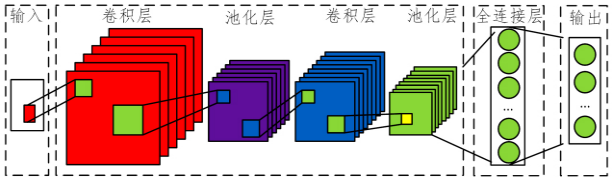


图 2 卷积神经网络(CNN)结构

Fig. 2 Convolutional neural network (CNN) structure

长短期记忆网络(LSTM)是递归神经网络(RNN)的改进方法，目的在于缓解爆炸梯度问题。与传统 RNN 单元相比，LSTM 使用一组 gate 函数控制反馈，这样短时间的错误最终会被删除，那些持久的特性会被保留下来。其数据处理流程如图 3 所示。

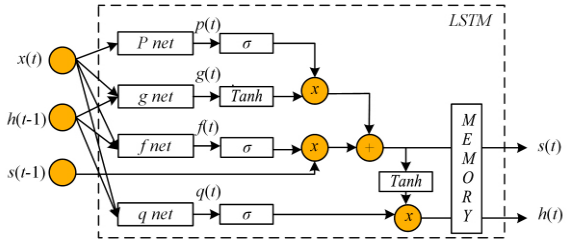


图 3 长短期记忆网络(LSTM)数据处理图

Fig. 3 Long and short term memory network (LSTM) data processing chart

LSTM 抽象为 4 个子网(即 p-net，g-net，f-net 和 q-net)、1 组门控制和 1 个内存组件的连接。图中的输入和输出是由输入 $x_t$ 决定的相同大小的向量。状态 $s(t)$ 保存当前学习的反馈。LSTM 中的所有子网都具有相似的结构，如式(1)所示：

$$b+U\times=x(t)+W\times h(t-1) \tag{1}$$

其中，$x(t)$，$h(t-1)$，$b$，$U$ 和 $W$ 分别为电流输入、前向输出、偏置、电流权值矩阵输入和前一个输出的递归权矩阵，4 个网中的每一个都有不同的 $b$，$U$ 和 $W$。然后，使用子网中的 $p(t)$，$g(t)$，$f(t)$ 和 $q(t)$ 进行输出，通过两种类型的控制门($\sigma$ 和 tanh)来确定先前学习的反馈 $s(t)$ 和电流输出 $h(t)$，具体如式(2)、式(3)所示：

$$s(t)=\sigma(f(t))*s(t-1)+\sigma(p(t))*\tanh g(t) \tag{2}$$

$$h_t=\tanh s(t)*\sigma(q(t)) \tag{3}$$

LSTM 通过调整网络中的权值和 $\sigma$ 值来学习输入，从而

在输出中有效地生成输入数据之间的时间特征。

CNN-LSTM 算法具有时间和空间的特征表达能力，由于入侵攻击的时间是持续的过程，因此攻击的手段是多种多样的，针对的攻击对象或攻击点也是不同的。CNN 用于提取特征，通过卷积核操作可以提取高层特征，这在图像处理领域已经有了成功的应用。LSTM 通过门函数控制历史数据的记忆和遗忘，适合于处理长时间序列数据，提高了检测准确率。因此，本文采用 CNN-LSTM 的算法模型进行入侵检测处理是合适的。CNN-LSTM 算法模型如图 4 所示，其具体步骤如下：

Step 1 输入层通过流量数据采集模块采集实时车联网数据，本文通过数据集进行特征分析，包括网络协议类型、网络服务类型、网络连接状态和连接时间等。

Step 2 按照数据处理的步骤，对数据分别进行预处理、数值化以及归一化操作。具体操作步骤后文将详细说明。

Step 3 将处理后的数据送入卷积层进行特征提取，通过一维卷积运算输出特征。每个卷积层之后伴随着一个池化层用于减少特征维度，加速收敛，去除冗余特征，防止网络过拟合。随后通过全连接层将所有局部特征进行整合形成整体特征。最后经过全连接层中的 LeakyReLU 激活函数进行操作。

Step 4 将 CNN 提取出的特征输入到 LSTM。经过 softmax 函数得到网络数据的分类结果。

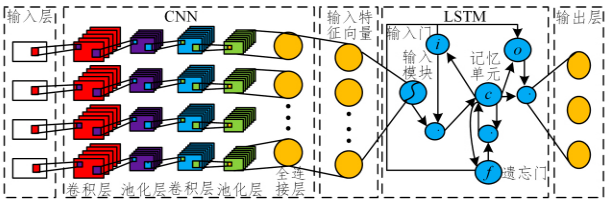


图 4 CNN-LSTM 算法结构示意图

Fig. 4 CNN-LSTM algorithm structure diagram

### 2.3 Spark 框架

为了提高检测效率，本文应用了 Apache Spark 框架，这是一个围绕速度、易用性和复杂分析构建的大数据处理框架，2009 年由加州大学伯克利分校开发[17,21]，2010 年成为了 Apache 开源项目之一，与 Hadoop 和 Storm 等其他大数据以及 MapReduce 技术相比，Spark 有如下优势：

(1)Spark 提供了一个全面、统一的框架用于管理各种不同性质(文本数据、图表数据等)的数据集和数据源(批量数据或实时的流数据)。

(2)Spark 将 Hadoop 集群中应用在内存中的运行速度提升 100 倍，在磁盘上的运行速度提升 10 倍。

(3)相比 MapReduce，Spark 数据计算能力更快且提供了更加丰富的功能。

当处理的数据量超过单机尺度(计算机有 4 GB 内存，需要处理 100 GB 以上的数据)或者处理的数据量不大，但计算很复杂，需要大量时间时，可以利用 Spark 集群的强大计算资源，并行化地进行计算。其架构示意如图 5 所示。

利用 Spark 分布式开源框架，将本次实验 PC 机连接，形成主从式的控制结构。主节点对从节点进行任务调度、分发和容错，从节点实现并行计算，此结构被证明是一个拥有高可靠、高并发、高性能计算能力的分布式结构。随后利用节点的 HDFS 存储系统对数据进行存储，并利用组合深度学习算法进行入侵检测。具体流程如图 6 所示。

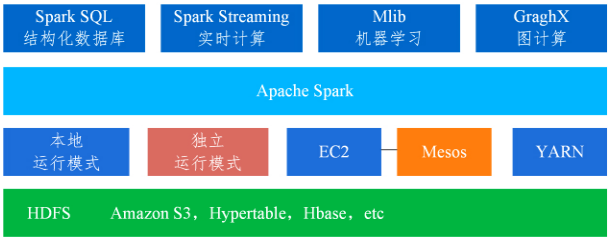


图 5 Apache Spark 架构

Fig. 5 Apache Spark architecture diagram

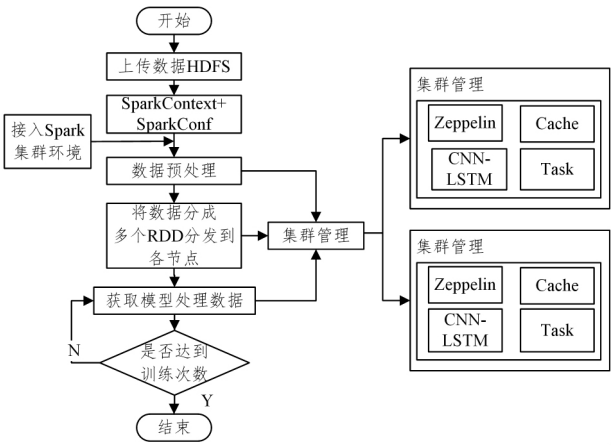


图 6 Spark 分布式组合深度学习算法流程图

Fig. 6 Flowchart of Spark distributed combined deep learning algorithm

## 3 实验分析

### 3.1 实验环境说明

在本次实验中，实验分布式平台采用 1 台主节点和 4 台从节点的物理主机搭建。其中，Spark 版本为 1.5.0，实验软件程序使用 Python3.7 进行编写，实现了 Spark 架构下的基于 CNN-LSTM 组合深度学习算法的车联网入侵检测方法(见图 1)。实验操作系统为 Ubuntu16.04。硬件条件为：主节点 Inter © Corei7-6500CPU @ 3.2GHz 处理器和 8.0GB RAM，从节点 Inter © Corei5-6500CPU@3.2GHz 处理器和 4.0GB RAM。在训练数据时，进行 1 000 次训练。

### 3.2 数据集说明

#### 3.2.1 NSL-KDD 数据集

相比传统网络，车联网由车辆节点自组织形成车与车(Vehicleto Vehicle，V2V)以及车与基础设施(Vehicleto Infrastructure，V2I)的异构通信网络[20]。其网络特点与传统网络相比，区别在于车辆高速行驶、信道快速衰落、多普勒效应严重、网络拓扑变化快等。但是，车联网在通信过程中遭受的攻击手段与传统网络中的攻击方式有许多的共同之处，例如后门攻击、拒绝服务等。因此，为了评估提出的基于 Spark 的车联网分布式组合深度学习入侵检测方法，将深度学习算法应用于入侵检测两个基准数据集，即 NSL-KDD[18] 和 UNSW-NB15[18]，构建一个在车联网外部通信中有效的入侵检测系统。实验数据共包括 219 473 条训练数据和 51 025 条测试数据。

NSL-KDD 数据集是对 KDD CUP-99 的改进[19]。它除

去了 CUP-99 数据集中的冗余，解决了分类器偏向于重复出现的记录的问题。与原始 KDD99 数据集相比，使用 NSL-KDD 数据集进行分类会产生相似或更好的准确性，其作为被公认用来进行入侵检测实验最好用的数据集之一，数据集的攻击分为 4 个类别。

(1)拒绝服务(DoS)：入侵者会向服务器发送大量恶意请求，导致机器的内存和计算资源太满或者忙碌而无法为合法流量提供服务，从而拒绝正常用户服务。

(2)用户到根(U2R)：作为一种攻击类型，攻击者尝试通过初始的正常用户访问权限获得管理员权限。

(3)远程到本地攻击(R2L)：攻击者想要通过网络将数据发送到计算机并欺诈性地获得对该计算机的访问权限。

(4)探测攻击(Probe)：扫描网络以获取用户设备详细信息和漏洞信息，这些信息以后可用于对漏洞或执行其他类型的攻击。

表 1 NSL-KDD 数据集的攻击分类

Table 1 Attack classification of NSL-KDD dataset

| 攻击类型 | 训练集 | 测试集 |
|---|---|---|
| Dos | 11656 | 7458 |
| U2R | 52 | 200 |
| R2L | 995 | 2754 |
| Probe | 45927 | 2421 |
| Normal | 67343 | 9711 |
| Total | 125973 | 22544 |

### 3.2.2 UNSW-NB15 数据集

UNSW-NB15 数据集是澳大利亚网络安全中心(ACCS)以混合的方式生成的数据集。该数据集共有 9 种攻击类型，涉及随着时间变化而捕获的正常流量的现实活动，如表 2 所列。

表 2 UNSW-NB15 数据集攻击类型描述

Table 2 UNSW-NB15 data set attack type description

| 攻击类型 | 描述 | 训练集 | 测试集 |
|---|---|---|---|
| 正常数(Normal) | 正常连接数据 | 56000 | 37000 |
| 模糊攻击(Fuzzers) | 一种攻击程序，攻击者通过向程序、操作系统或网络输入大量的随机数据，使其崩溃，从而尝试发现安全漏洞 | 18184 | 6062 |
| 分析攻击(Analysis) | 一种通过端口(例如端口扫描)、电子邮件(例如垃圾邮件)和网络脚本(例如 HTML 文件)渗透到 Web 应用程序的各种入侵 | 2000 | 677 |
| 后门攻击(Backdoors) | 一种绕过秘密常规身份验证，保护对设备的未经授权的远程访问以及在纯文本正难以继续运行时定位其入口的技术 | 1746 | 583 |
| 拒绝服务(Dos) | 一种入侵行为，它通过内存破坏计算机资源，变得非常繁忙，以阻止授权的请求访问设备 | 12264 | 4089 |
| 漏洞利用(Exploits) | 利用主机，网络上有意或无意的行为所引起的故障，错误或漏洞的一系列指令 | 33393 | 11132 |
| 泛攻击(Generic) | 一种使用散列函数对每个块密码建立冲突的技术，而无需考虑块密码的配置 | 40000 | 18871 |
| 侦察攻击(Reconnaissance) | 可以定义为探测：收集有关计算机网络信息以逃避其安全控制的攻击 | 10491 | 3496 |
| 缓冲区溢出(Shellcode) | 一种攻击程序，攻击者从外壳程序开始穿透一小段代码来控制受感染的计算机 | 1133 | 378 |
| 蠕虫攻击(Worms) | 攻击者通过复制自身以在其他计算机上传播的攻击。通常，它使用计算机网络进行传播，具体取决于访问它的目标计算机上的安全故障 | 130 | 44 |
| 合计 | | 93500 | 28481 |

此外，该数据集包含 49 种特征，这些特征组成了存在于主机和网络数据包之间的流量，用于区分正常或异常的观测结果。与其他数据集相比，它包含真实场景数据和合成攻击行为，另外 UNSW-NB15 的复杂性意味着这个数据集是有效可靠的。

## 3.3 数据处理

### 3.3.1 数据预处理

数据预处理的主要目的是对数据集进行规范化和数值化，具体来说就是对训练集和测试集进行检测和处理。例如，清洗错误数据或丢失不全的数据，对没有数值化的数据进行数值化，最后进行数据的分组和转化以获得有价值的新数据。数据预处理包括数值化和归一化操作，如图 7 所示。

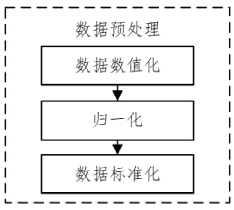


图 7 数据预处理操作

Fig. 7 Data preprocessing operation

### 3.3.2 数据数值化

由于 CNN 模型只能识别数值型数据，而原始的 NSL-KDD 和 UNSW-NB15 数据集中包含了字符型数值，因此，为了能够更好地分析和识别数据集中的内容，需要对数据集进行数值化操作。例如，属性特征 protocol_type 有 3 种协议类型，即 TCP，UDP，ICMP，将其编码为 1，2，3。

### 3.3.3 归一化

归一化的目的一是把数据变成(0，1)或者(1，1)之间的小数，以更加方便快速的把数据提取出来；二是数据经过数值化操作之后，数值化之后不同的量纲和量纲单位会影响到数据分析的结果，为了消除指标之间的量纲影响，需要进行归一化操作。归一化常用的两种方法如下：

(1)离差标准化。对数据进行线性变换，结果值映射到[0，1]。其函数公式为：

$$X^* = \frac{X - min}{max - min} \tag{4}$$

其中，$max$ 为数据的最大值，$min$ 为数据的最小值。

(2)0 均值标准化。该方法把数据的均值和标准差进行数据的标准化，经过处理的数据符合标准正态分布，即均值为 0，标准差为 1，函数原型为：

$$X^* = \frac{X-\mu}{\sigma} \tag{5}$$

其中，$\mu$ 为所有数据的均值，$\sigma$ 为所有数据的标准差。本文选用 0 均值标准化对数据进行归一化处理得到标准数据。

### 3.4 实验结果分析

本文以准确率(AC)、误报率(FPR)以及检测时间为评价整套车联网入侵检测性能的关键指标。准确率是衡量正确预测攻击和非攻击的正常流量的能力。其指标的详细定义如式(6)、式(7)所示：

$$AC=\frac{TP+TN}{TP+TN+FP+FN} \tag{6}$$

$$FPR=\frac{FP}{FP+TN} \tag{7}$$

表 3 列出了混淆矩阵。其中，TP 和 TN 分别为被正确分类的攻击次数和正常流量次数；FP 是实际正常记录的数量被误分类为攻击，而 FN 是误分类为正常流量的攻击数量。

表 3 混淆矩阵

Table 3 Confusion matrix

| 预测值 | 真实值 | |
|---|---|---|
| | 攻击类 | 正常类 |
| 攻击类 | TP | FN |
| 正常类 | FP | TN |

#### 3.4.1 与其他深度学习的检测性能对比分析

将 CNN-LSTM 算法与 SVM、RNN、CNN 和 LSTM 算法对不同的攻击类型进行准确率(AC)和误报率(FPR)的对比分析，实验结果如表 4 所列。CNN-LSTM 方法在分类检测率方面高于其他算法，且有较低的误报率。

表 4 CNN-LSTM 数据集不同攻击类型的性能比较

Table 4 Performance comparison of different attack types of CNN-LSTM dataset

| 攻击类型 | SVM | | RNN | | CNN | | LSTM | | CNN-LSTM | |
|---|---|---|---|---|---|---|---|---|---|---|
| | AC | FPR | AC | FPR | AC | FPR | AC | FPR | AC | FPR |
| Normal | 96.38 | 2.86 | 98.78 | 2.10 | 99.70 | 1.40 | 98.30 | 2.17 | 99.80 | 2.80 |
| DoS | 95.49 | 4.72 | 98.98 | 3.20 | 83.10 | 4.90 | 82.50 | 2.40 | 97.40 | 0.80 |
| R2L | 62.49 | 3.78 | 73.55 | 4.31 | 63.30 | 0.70 | 68.45 | 2.12 | 78.40 | 0.01 |
| U2L | 13.20 | 8.50 | 28.53 | 5.74 | 32.51 | 0.00 | 42.44 | 6.49 | 67.80 | 0.00 |
| Probe | 93.57 | 2.77 | 93.21 | 3.16 | 71.20 | 2.30 | 80.36 | 3.54 | 94.10 | 0.50 |
| Fuzzers | 98.90 | 0.00 | 97.20 | 2.10 | 96.50 | 2.10 | 96.62 | 2.30 | 99.00 | 0.00 |
| Analysis | 97.60 | 0.00 | 96.30 | 2.60 | 97.80 | 0.00 | 95.81 | 0.00 | 98.10 | 3.10 |
| Backdoors | 95.80 | 0.00 | 96.70 | 0.00 | 95.30 | 0.00 | 94.30 | 0.00 | 96.20 | 0.00 |
| Exploits | 89.50 | 0.00 | 90.10 | 0.00 | 90.60 | 2.10 | 87.40 | 1.10 | 91.30 | 0.00 |
| Generic | 80.60 | 2.40 | 74.40 | 0.00 | 78.10 | 0.00 | 75.30 | 0.00 | 86.40 | 1.20 |
| Reconnaissance | 91.60 | 3.90 | 89.50 | 4.10 | 88.60 | 1.30 | 87.00 | 2.10 | 92.00 | 3.40 |
| Shellcode | 93.00 | 0.00 | 91.60 | 0.00 | 92.80 | 0.00 | 90.90 | 0.00 | 94.70 | 0.00 |
| Worms | 91.80 | 0.00 | 92.00 | 0.00 | 91.10 | 0.00 | 90.20 | 0.00 | 93.80 | 0.00 |

为了验证实验整体的有效性和对比性，本文利用 NSL-KDD 和 UNSW-NB15 两个数据集对上述 5 种算法的准确率(AC)和误报率(FPR)进行对比实验分析。实验结果如图 8 和图 9 所示。

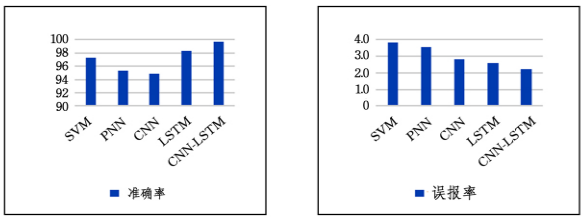


图 8 NSL-KDD 数据集下各算法性能比较

Fig. 8 Performance comparison of algorithms in NSL-KDD dataset

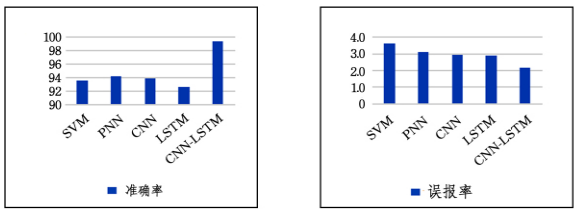


图 9 UNSW-NB15 数据集下各算法性能比较

Fig. 9 Performance comparison of algorithms in UNSW-NB15 dataset

通过图 8 和图 9 可以看出，CNN-LSTM 在 NSL-KDD 数据集和 UNSW-NB15 数据集中表现良好，分别达到了 99.7%和 99.4%的准确率，同样也拥有最低的误报率，分别为2.24%和 2.17%。因此，该算法在同类算法中拥有较好的性能特性。

#### 3.4.2 与其他非分布式深度学习的时间对比分析

本文所有的深度学习算法都在 Apache Spark 下以分布式方式实现，实验结果如图 10 和图 11 所示。可以看到，与传统的非并行机器和深度学习算法相比，训练和测试的时间大大缩短。实验结果表明，CNN-LSTM 算法所用的训练时间和测试时间最短。

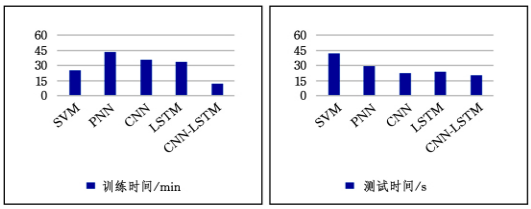


图 10 NSL-KDD 数据集下各算法检测时间对比

Fig. 10 Comparison of detection time of each algorithm in NSL-KDD dataset

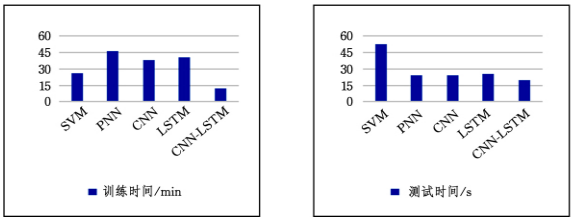


图 11 UNSW-NB15 数据集下各算法检测时间对比

Fig. 11 Comparison of detection time of each algorithm in UNSW-NB15 dataset

**结束语** 通过对比实验发现，针对入侵检测系统存在的面向大数据检测速度慢、检测效率低的问题，充分考虑分布式框架和深度学习算法的优势，将分布式架构与深度学习 CNN-LSTM 算法相结合。通过数据预处理等方式对数据进行标准化的处理之后，提高了检测效率和检测时间。在 NSL-KDD 数据集和 UNSW-NB15 数据集上进行实验验证，结果表明，使用 Spark 框架的 CNN-LSTM 的深度学习算法与其他深度学习算法相比，降低了训练时间和测试时间，提高了检测率，可以很好地满足入侵检测实时性的要求，更加满足车联网对入侵检测的实际需要。

在下一步工作中，在本文提高入侵检测性能、降低检测时间的基础上，进一步围绕深度学习算法的检测能力，在分布式平台上进行入侵检测，探索合适的分布式深度学习算法来满足车联网信息安全入侵检测的需要。同时考虑更加高效的算法处理车联网的网络数据流量，增强算法的适应性。

## 参 考 文 献


[1] HALIMAA A A, SUNDARAKANTHAM K. Machine learning based intrusion detection system[C]// Proceedings of the International Conference on Trends in Electronics and Informatics (ICOEI 2019). 2019: 916-920.

[2] VINAYAKUMAR R, SOMAN K P, POORNACHANDRANY P. Applying convolutional neural network for network intrusion detection[C]// 2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI 2017). 2017: 1222-1228.

[3] DONG B, WANG X. Comparison deep learning method to traditional methods using for network intrusion detection[C]// Proceedings of 2016 8th IEEE International Conference on Communication Software and Networks (ICCSN 2016). IEEE, 2016: 581-585.

[4] ISHAQUE M, HUDEC L. Feature extraction using Deep Learning for Intrusion Detection System[C]// 2nd International Conference on Computer Applications and Information Security (ICCAIS 2019). IEEE, 2019: 1-5.

[5] YAO Y, YANG W, GAO F X, et al. Anomaly intrusion detection approach using hybrid MLP/CNN neural network[J]. Proceedings-ISDA 2006: Sixth International Conference on Intelligent Systems Design and Applications, 2006, 2 (60473073): 1095-1102.

[6] PENG W, KONG X, PENG G, et al. Network intrusion detection based on deep learning[C]// Proceedings-2019 International Conference on Communications, Information System, and Computer Engineering (CISCE 2019). IEEE, 2019: 431-435.

[7] DING W H, ZHOU K, LONG Y Y, et al. Research on Intrusion Detection Based on deep convolution neural network[J]. Computer Science, 2019(10): 1-11.

[8] CHOCKWANICH N, VISOOTTIVISETH V. Intrusion Detection by Deep Learning with TensorFlow[J]. International Conference on Advanced Communication Technology, ICACT, Global IT Research Institute (GIRI), 2019(2): 654-659.

[9] DOBSON A, ROY K, YUAN X, et al. Performance Evaluation of Machine Learning Algorithms in Apache Spark for Intrusion Detection[C]// 2018 28th International Telecommunication Networks and Applications Conference (ITNAC 2018). IEEE, 2019: 1-6.

[10] MENG X, BRADLEY J, YAVUZ B, et al. MLlib: machine leaning in apache spark[J]. Computer Science, 2015, 17(1): 1235-1241.

[11] CHEUNG L. The rise and predominance of Apache Spark," infoworld. com [OL]. https://www. infoworld. com/article/3216144/spark/ apache-spark. html.

[12] HOCHREITER SSCHMIDHUBER J. Long short-term memory [J]. Neural Computation, 1997, 9(8): 1735-1780.

[13] REVATHI S, MALATHI A. A detailed analysis on NSL-KDD dataset using various machine learning techniques for Intrusion detection[J]. ESRSA Publications, 2013.

[14] MOUSTAFA N, SLAY J. Unsw-nb15: a comprehensive data set for network intrusion detection systems (unsw-nb15 network data set)[C]// 2015 Military Communications and Information Systems Conference (MilCIS). IEEE, 2015: 1-6.

[15] ZHOU F Y, JIN L P, DONG J. Summary of Research on Convolutional Neural Networks[J]. Chinese Journal of Computers, 2017, 40(6): 1229-1251.

[16] ANZER A, ELHADEF M. A multilayer perceptron-based distributed intrusion detection system for internet of vehicles[C]// Proceedings-4th IEEE International Conference on Collaboration and Internet Computing (CIC 2018). IEEE, 2018: 438-445.

[17] VIMALKUMAR K, RADHIKA N. A big data framework for intrusion detection in smart grids using apache spark[C]// 2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI 2017). 2017: 198-204.

[18] FERRAG M A, MAGLARAS L, MOSCHOYIANNIS S, et al. Deep learning for cyber security intrusion detection: Approaches, datasets, and comparative study[J]. Information Security Technical Report, 2020, 50(2): 102419. 1-102419. 19.

[19] RING M, WUNDERLICH S, SCHEURING D, et al. A survey of network-based intrusion detection data sets[J]. Computers and Security, Elsevier Ltd, 2019, 86: 147-167.

[20] SEDJELMACI H, SENOUCI S M, ABU-RGHEFF M A. An efficient and lightweight intrusion detection mechanism for service-oriented vehicular networks[J]. IEEE Internet of Things Journal, IEEE, 2014, 1(6): 570-577.

[21] GAO Y, WU H, SONG B, et al. A distributed network intrusion detection system for distributed denial of service attacks in vehicular ad hoc network[J]. IEEE Access, 2019, 7: 154560-154571.


**YU Jian-ye**, born in 1993, postgraduate. His main research interests include information security.

**QI Yong**, born in 1970, Ph.D, professor, Ph.D supervisor, is a member of China Computer Federation. His main research interests include traffic bigdata, security of internet of vehicles.