计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2019,28(8):46−52 [doi: 10.15888/j.cnki.csa.006979]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于 Spark 的电网工控系统流量异常检测平台①

张艳升[1,2], 李喜旺[1], 李锦程[3]

[1](中国科学院 沈阳计算技术研究所, 沈阳 110168)
[2](中国科学院大学, 北京 100049)
[3](国网辽宁省电力有限公司, 沈阳 110004)
通讯作者: 张艳升, E-mail: 17615850921@163.com

摘　要: 针对传统的电力网络流量检测安全预警系统在面对海量高维度数据时, 其在精度、实时性、扩展性以及效率上都无法满足需求的问题, 建立出一种基于 Spark 的电网工控系统流量异常检测平台. 该平台以 Spark 为计算框架, 主要由数据采集与网络流量深度包检测协议解析模块, 实时计算数据分析处理模块, 安全预警预测模块和数据存储模块组成, 为流量异常检测提出了一套完整的流程. 实验结果表明, 该平台能够有效地检测出异常流量, 做出安全预警, 方便工作人员及时做出决策, 这充分说明该平台非常适用于电力控制系统, 能够应对海量高维复杂数据做出实时分析以及安全预警, 极大地提高了电网工控系统的安全性能.

关键词: Spark; 流量异常检测; 电网工控系统; Kafka; Deep Learing 4J

引用格式: 张艳升,李喜旺,李锦程.基于 Spark 的电网工控系统流量异常检测平台.计算机系统应用,2019,28(8):46–52. http://www.c-s-a.org.cn/1003-3254/6979.html

Flow Anomaly Detection Platform for Power Grid Industrial Control System Based on Spark

ZHANG Yan-Sheng[1,2], LI Xi-Wang[1], LI Jin-Cheng[3]

[1](Shenyang Institute of Computing Technology, Chinese Academy of Sciences, Shenyang 110168, China)
[2](University of Chinese Academy of Sciences, Beijing 100049, China)
[3](State Grid Liaoning Electric Power Co. Ltd., Shenyang 110004, China)

Abstract: Aiming at the problem that the traditional power network traffic detection and security warning system cannot meet the demand in terms of accuracy, timeliness, expansibility, and efficiency in facing of massive high-dimensional data, a Spark based traffic anomaly detection platform for power grid industrial control system is established. The platform takes Spark as its computing framework, which is mainly composed of data acquisition and network traffic deep packet detection protocol parsing module, real-time computing data analysis and processing module, security warning and prediction module, and data storage module, to complete process for traffic anomaly detection. Experimental results show that the platform can effectively detect the abnormal flow, make the safety warning, convenient for staff to make decisions in time. This fully shows that the platform is very suitable for electric control system, can deal with massive amounts of high-dimensional complex data real time analysis and early warning, greatly improve the safety performance of the power grid control system.

Key words: Spark; flow anomaly detection; power grid industrial control system; Kafka; Deep Learning 4J

① 基金项目: 国家科技重大专项 (2017ZX01030-201)
Foundation item: National Science and Technology Major Program (2017ZX01030-201)
收稿时间: 2019-01-14; 修改时间: 2019-02-03; 采用时间: 2019-02-18; csa 在线出版时间: 2019-08-08

两化融合的力度深度加强, 在电网领域越来越多地应用的信息技术. 工业控制系统已广泛应用于电力工业领域. 其中, 在电网领域很大一部分的基础设施需要依靠工业控制系统来实现自动化作业. 电网工控系统通常情况下采用通用的软硬件和网络设施, 并越来越广泛地与企业网和互联网等集成, 形成了开放的网络环境. 由于传统电网控制系统是基于物理隔离的, 主要关注系统的功能安全, 缺乏对网络信息安全的考虑, 没有专门的安全防御措施. 例如, 2010 年震撼全球的“震网”病毒事件, 专门攻击工业控制系统设施, 造成伊朗核电站推迟发电[1,2]. 电力控制系统网络化的快速发展, 相应导致了系统的安全风险不断增加, 面临的网络安全问题更加突出, 再加上利用 TCP/IP 技术对工业网络进行通信, 对 TCP/IP 进行相关攻击, 能够快速进入网络, 使得工控系统面临更大的安全挑战. 大数据技术的发展迅速, 利用大数据技术来解决工业控制系统中异常流量检测问题是当今研究的新的方向, 如果当人力无法处理超出可控范围内的异常流量数据时, 这样一些隐藏的网络攻击将有机可趁, 这时利用大数据技术来处理大规模数据中的异常情况就非常有必要, 相关工作人员能够快速定位与分析, 做出相应地措施.

随着电力工控系统智能化建设的不断发展, 数据呈现出海量化、高维化、复杂化的趋势. 传统的网络监测安全预警系统在面对海量高维度数据时, 其在精度、实时性、扩展性以及效率上都无法满足需求. 目前大多数的工业控制网络安全预警方案主要以 Hadoop 数据处理平台为载体, 考虑在集群上部署 Hadoop 数据处理平台, 然后通过使用统计学习方法或者各种机器学习算法对采集到的海量数据进行建模, 从而进行大规模的离线计算, 然而利用 Hadoop 为载体的系统并没有考虑到预警预测的实时性, 也没有考虑到其处理的速度问题, 不能快速、实时的反馈安全预警, 利用 Hadoop 为载体的机器学习方法也无法做到在内存中大量迭代. 要想快速、实时、高效的对网络做出安全预警, 需要研究其他大数据处理技术, 例如, 近几年出现的新型的数据处理计算框架 Spark, 它作为一种基于内存的编程模型, 它将迭代过程和中间结果放在内存中进行, 数据处理速度上得到很大提升[3,4]. 它的组件 Spark Streaming 能够对数据流进行实时处理, 从而能够满足实时性的要求[5−7], 但是它们都是利用 Spark 内置的简单的机器学习模型来进行建模预测, 在面对海量数据集时无法体现其优势, 本文在利用 Spark 计算框架的基础上, 结合与其相集成的 Deep Learning 4J 来进行深度学习模型的建模, 这样能够统一技术栈, 不论在精度、实时性还是扩展性、效率上都有进一步提升.

因此, 在大数据环境下, 为了能够对工业控制网络安全作出快速、精确和实时的预警, 考虑使用 Spark 计算框架, 结合利用深度学习建模, 充分利用其低延迟、可扩展以及高可用的特性, 这显然是发展的一种趋势.

1 电网工控系统流量异常检测系统的设计

基于 Spark 的工业控制网络安全预警平台核心技术模块主要包括数据采集与网络流量深度包检测协议解析模块、实时数据分析处理模块、安全预警预测模块和数据存储模块共计 4 个系统功能模块. 系统的整体结构图如图 1 所示.

1.1 数据采集与网络流量深度包检测协议解析模块

数据采集层负责从不同的数据源采集数据. 利用网络流量深度包检测传感器通过旁路接入的方式实现对网络内数据的检测. 深度包检测技术 (Deep Packet Inspection, DPI) 是在传统 IP 数据包检测技术 (OSI L2-L4 之间包含的数据包元素的检测分析) 之上增加了对应用层数据的应用协议识别, 数据包内容检测与深度解码. 一般网络设备只会查看以太网头部、IP 头部而不会分析 TCP/UDP 里面的内容这种被称为浅数据包检测; 与之对应的 DPI 会检查 TCP/UDP 里面的内容, DPI 数据包检测如图 2 所示.

利用 Kafka(是一种高吞吐量的分布式发布订阅消息系统) 作为平台数据的接入与分发. Kafka 能够灵活的处理流式数据, 主要为本文系统平台中的数据采集层与数据预处理层之间提供高性能与低延迟的数据流转[8]. 一个典型的 Kafka 体系架构如图 3 所示, 主要包括若干 Producer, 若干 broker, 若干 Consumer (Group), 以及一个 Zookeeper 集群. Kafka 通过 Zookeeper 管理集群配置, 选举 leader, 以及在 consumer group 发生变化时进行 rebalance. 将数据采集层采集好的特征数据根据网络流量深度包传感器接入不同设备来划分 Kafka 中不同的 Topic, 每一 Topic 又划分多个 partition, 然后 Kafka 中的 Producer 使用 push(推) 模式将消息发布到 broker, 发送每一条数据到 broker, Producer 会根

据这条消息的 Key 和 partition 机制来决定发到哪一个 partition, 数据处理层利用 Consumer 使用 pull(拉) 模式从 broker 中订阅并根据 Topic 主题来消费消息.

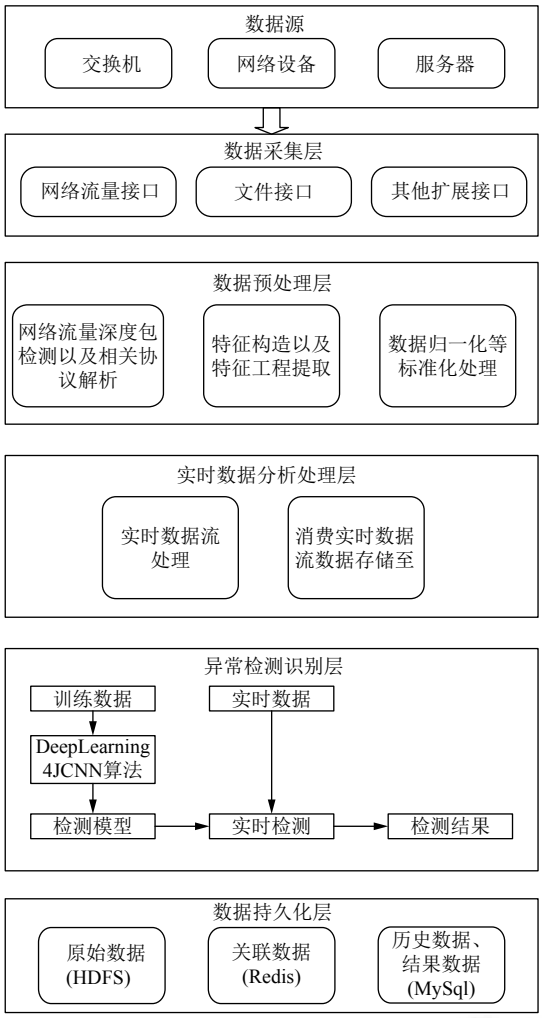

图 1 异常流量检测系统整体架构图

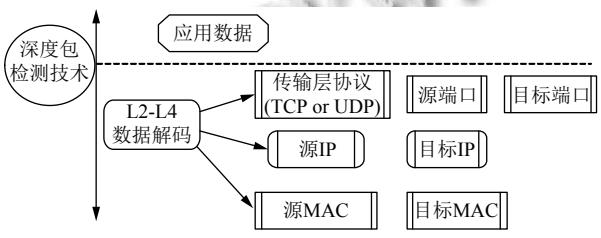

图 2 DPI 数据包检测图

1.2 实时计算数据分析处理模块

该模块主要建立以 Spark 为核心的实时数据流处理模块, 主要面对大量实时监测的数据. 采用 Spark Streaming(是建立在 Spark 上的实时计算框架, 通过它提供丰富的 API、基于内存的高速执行引擎, 用户可以结合流式、批处理和交互式查询应用) 技术进行实时事件流的处理, 它将流式计算分解成一系列连续的 DStream(Discretized Stream, DStream)[9–11]. DStream 实质上 RDDs 的集合, 主要是以时间为键, RDD 为值的哈希表, 保存以时间为顺序产生的 RDD, 而每个 RDD 封装了批处理时间间隔内获取的数据, 每次产生的 RDD 会被添加到哈希表中, 如果不需要的 RDD 会从哈希表中删除. 该模块主要消费来自 Kafka 中的数据, 利用 Spark Streaming 的 Direct 方式来消费来自 Kafka 中数据, 这种方式能够定期的查询 Kafka 中的 topic+partition 中的偏移量, DStream 会创建与 Kafka 分区一样的 RDD 分区数, 能够并行的从 Kafka 中消费数据. Spark Streaming 从 Kafka 中消费数据运行原理图如图 4 所示. 该模块一方面消费 Kafka 中的数据, 可以将处理后的数据存储至历史数据库, 方便数据分析以及 Spark ML 建模使用; 另一方面, Spark Streaming 可以处理实时监测的数据, 可以存入内存数据库 Redis 实时展示网络流量动态曲线, 同时实时处理的数据输入到已建立好的流量异常检测模型中来检测, 这样方便在出现异常时能够实现及时预警, 方便工作人员作出及时调整.

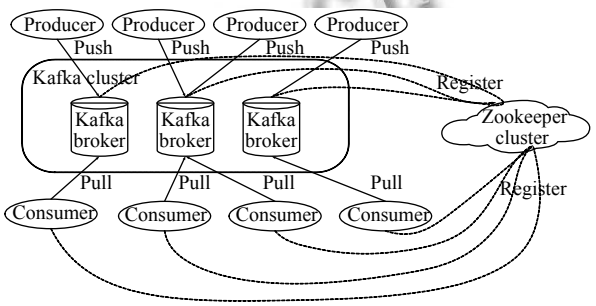

图 3 Kafka 体系架构图

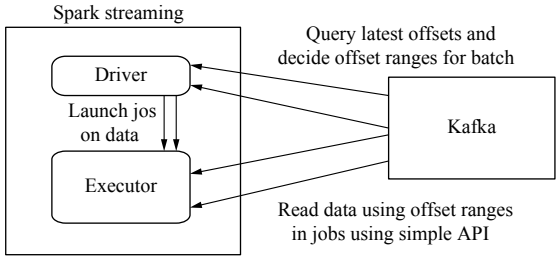

图 4 Spark Streaming 消费 Kafka 原理图

Spark Streaming 消费 Kafka 中的数据, 首先识别原始数据中的每条特征是否是字符形式的数据, 如果是, 将该特征进行数值化处理, 最终每条数据都是数值化的形式, 然后将转化为数值化后的数据进行最大值最小值标准化处理, 方便接下来利用 Deep Learning 4J 设计的卷积神经网络的使用, 将其一维转二维化处理.

1.3 安全预警预测模块

电网工控系统数据海量化以及复杂化, 对异常流量检测安全预警需要要求准确性, 实时性等要求, 通过读取历史数据库中的数据, 利用 Deep Learning 4J 建立以卷积神经网络为基础的流量异常检测模型, Deep Learning 4J 集成了 Hadoop 及 Spark, 能够与该平台实现统一的技术栈, 实现统一的 pipeline, 节省了研发成本[12–14]. 该模块能够根据历史数据库实现模型的建立, 每过一段时间可以增量更新模型, 同时根据 Spark Streaming 实时处理的数据, 利用该模型进行检测, 能够做出极高的准确率, 极大地提高了异常流量的效率以及精度.

1.4 数据存储模块

该模块主要根据数据的类型以及使用的情况进行选择, 数据采集的原始数据可以存放至 HDFS, 方便 Spark 快速读取, 以及历史数据挖掘分析. 为提高平台数据处理分析以及异常流量检测的实时性, 使用分布式内存数据库 Redis, 主要利用 Redis 数据库的键值存储, 对实时数据处理的结果进行存储. 为方便利用 Deep Learning 4J 中的卷积神经网络建立模型, 将平台的离线历史数据存储用 Mysql 进行存储, 同时将异常流量检测模型检测的结果数据也存储至 Mysql 中, 这样可以积累训练数据, 同时满足增量更新的要求, 更有利于提高模型的准确性.

2 基于 Spark 的异常检测模型

电网工控系统中流量异常检测是非常重要有必要的, 网络运行安全时网络各个维度的特征都比较平稳. 但当异常发生时会产生较大的波动, 当某些维度的数值超过阈值时判定异常发生. 异常有许多种, 并且一种异常发生时对应的特征波动情况也相对稳定, 本文中主要对电网工控系统采集流量数据, 分为正常流量和部分少量异常流量, 采用 Spark 框架对数据进行处理, 为结合 Spark 以及达到实时性已经准确性的要求, 建立模型部分采用与 Spark 计算引擎集成的 Deep Learning 4J, 设计好卷积神经网络的网络结构, 这样能够很好捕捉好异常的波动, 达到较高的识别率. 异常检测模型如图 5 所示, 基于 Spark 的异常检测模型的算法如下:

1) 首先采集电网工控系统的数据, 将字符型数据转化为数值型, 并将数据进行标准化的形式处理.

2) 将每条数据即一维特征进行二维化处理, 将每条样本数据转化为图片的形式.

3) 设计卷积神经网络结构, 有卷积层, 池化层, Dropout 层, 全连接层以及反卷积层实现.

4) 利用设计好的卷积神经网络对训练数据集进行训练, 并利用测试样本对模型进行调参, 使模型达到最优.

5) Spark Streaming 实时处理的监测数据用训练好的模型进行预测, 从而作出识别预警, 并将识别结果存储至结果数据库中.

6) 定期增量更新模型, 方便提高系统的性能.

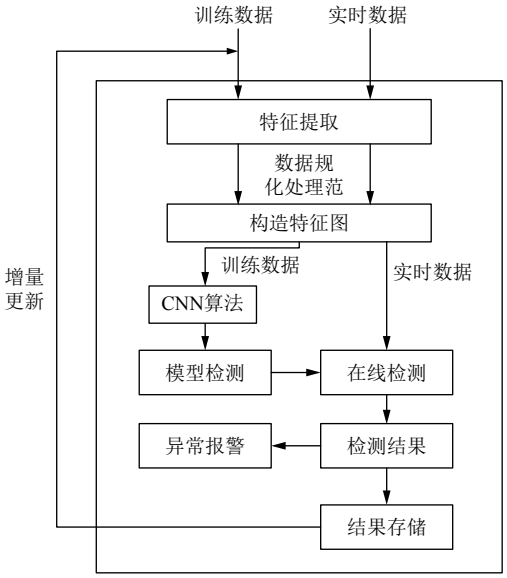

图 5 Deep Learning 4J 异常检测模型

卷积神经网络基本结构有输入层、卷积层、池化层、全连接层和输出层这几部分交替组合而成. 设计的卷积神经网络如图 6 所示.

本文中在卷积神经网络模型设计中加入了 dropout 技术, 其目的是为了提高模型的泛化能力, 进而能够对模型的过拟合现象得到有效的改善. 在实验环节, 对不加入 dropout 技术与加入 dropout 技术两个

模型做了相关对比, 在模型设计与测试时, 根据经验值将 dropout 设置成 0.4, 统计模型在最后的损失与准确率趋于稳定的结果, 如表 1 所示.

表 1 可知, 不带 dropout 技术的模型的训练时的损失值要小于加入 dropout 技术的损失值, 由此可知, 在不加入 dropout 技术时, 当训练模型时, 模型为了最小化损失函数, 从而对训练数据过分拟合, 使得模型的损失值最小. 但加入 dropout 技术的模型的准确率要大于不加入 dropout 技术的准确率, 因为加入 dropout 的模型在模型参数更新时会随机选择一部分参数进行更新, 避免对固有模型参数的学习记录. 即模型再每一次迭代过程中, 都会生成一个个的小网络, 每次只针对此次生成的小网络进行参数学习更新, 各个神经元之间不会形成特定的组合对模型学习的影响, 使得参数的分布趋于均匀. 因此, 使得模型在测试数据集的表现效果更佳.

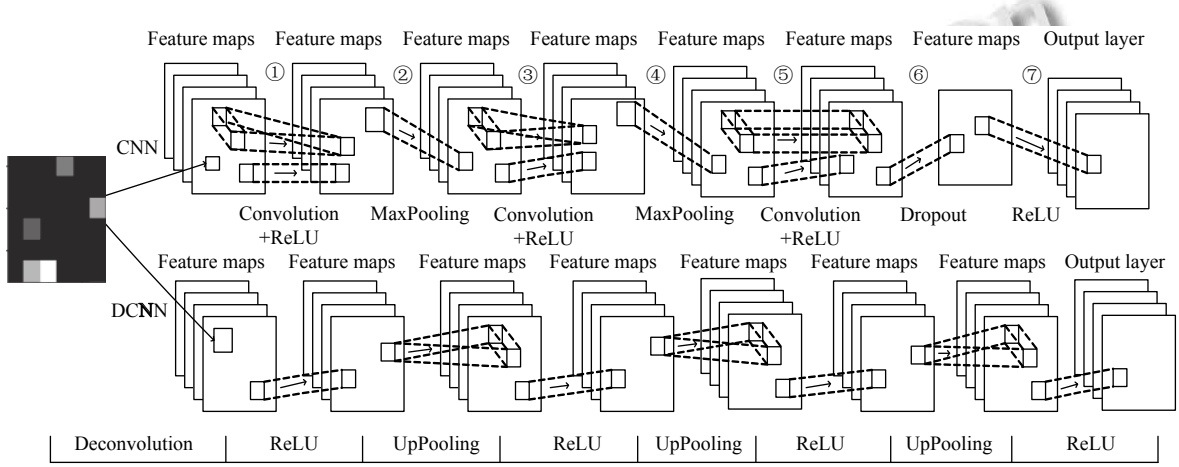

图 6 DeepLearning 4J 卷积神经网络结构图

表 1 Dropout 的影响

类别	模型损失值	模型准确率
去掉 dropout	0.21	0.91
加入 dropout	0.27	0.98

本文在卷积神经网络模型设计中加入了反卷积操作, 其反卷积操作的关键是卷积层的逆过程, 是能够对原始输入图像的重建, 从而训练得到参数更新, 能够降低模型的损失误差, 增加模型的泛化能力. 如表 2, 为模型在趋于稳定的时候统计的模型的损失与准确率的平均值. 从表 2 也可以看出加入反卷积层的模型的损失值更小, 泛化能力更强, 识别准确率更高.

表 2 反卷积层的影响

类别	模型损失值	模型准确率
不加入反卷积	0.32	0.91
加入反卷积	0.21	0.97

3 实验与结果分析

3.1 实验环境

平台的实验环境是基于分布式系统基础架构 Hadoop 安装的 Spark 集群, 集群中共有 10 机器作为节点, 是基于 Hadoop 中资源调度器 Yarn 来部署 Spark 为 Yarn Cluster 模式, 每台机器的内存为 32 GB. 集群软件配置如表 3.

表 3 集群软件配置

软件	版本
Hadoop	2.7.3
Spark	2.3.0
Kafka	2.0.0
Scala	2.12.8
Mysql	8.0.13
Redis	5.0.3
DeepLearning4J	1.0.0

3.2 实验数据与结果分析

本文所用数据是通过对东北电网采用网络流量深度包检测, 将检测的数据包进行捕获解析, 将解析的数据生产到 Kafka 相应话题, 利用 Spark Streaming 的时间窗的流式处理对数据进行处理, 采集网络数据时主要关注网络间与连接特征相关的信息. 例如, 在 TCP 连接的基本特征中可以提取连接的持续时间、网

络协议的类型、目标主机的网络服务类型、从源主机到目标主机的数据字节数等, 目前采集的网络流量数据集大小约为 4.62 GB.

在数据集的每条数据中都记录攻击类型, 其中攻击类型一共包括 4 个大类以及 28 个小类, 为了方便建模将数据集分为训练集与测试集, 28 个小类攻击类型中的 12 个小类出现在测试集中, 这样可以检验模型的泛化能力.

采集的流量数据样本特征值并不都是在 0~255 之间, 有些特征还是以字符串的形式采集下来的. 拿其中一条样本来看, 0, tcp, smtp, SF, 787, 329, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 0.00, 0.00, 0.00, 0.00, 1.00, 0.00, 0.00, 76, 117, 0.49, 0.08, 0.01, 0.02, 0.00, 0.00, 0.00, 0.00, normal.

首先将字符特征转化成数值型特征, 分别将 Internet 控制报文协议 (Internet Control Message Protocol, ICMP)、用户数据包协议 (User Datagram Protocol, UDP) 和 Tcp 3 种协议类型, 简单邮件传输协议 (Simple Mail Transfer Protocol, SMTP)、目标主机服务类型 ecr_i、private、超文本传输协议 (Hyper Text Transfer Protocol, HTTP) 等 70 种网络服务类型, SF(Symbol Flag)、REJ(REJECT) 等 11 种网络连接状态以及 Smurf 攻击 (Smurf)、Ping 扫射 (ping-sweep)、端口扫描 (port-scan)23 种攻击类型和一种正常状态转化数字标识. 对每一条转化好的数值特征采用数据标准化处理, 本文采用最大最小值归一化处理, 取每条数据的最大值最小值, 然后对每一个数据进行处理, 公式如下:

$$x^{*}=\frac{x-x_{\min}}{x_{\max}-x_{\min}} \tag{1}$$

将数据归一化的数据生成特征图的方式, 利用设计好的异常检测模型, 利用训练集与测试集来训练模型以及调优模型, 对 Spark Streaming 消费 Kafka 中的实时数据, 对数据进行相应处理, 利用训练好的模型进行相关预测, 以便做出安全预警.

检测结果主要利用检测率, 误报率和未知攻击检测率三个指标, 公式如下:

$$\begin{cases} \text{检测率}=\dfrac{\text{识别出的异常样本数}}{\text{所有异常样本数}}\times 100\% \\ \text{误报率}=\dfrac{\text{误判为异常的正常样本数}}{\text{所有正常样本数}}\times 100\% \\ \text{未知攻击检测率}=\dfrac{\text{识别出的未知攻击样本数}}{\text{所有未知攻击样本数}}\times 100\% \end{cases} \tag{2}$$

由表 4 列举出了 3 种攻击类型和整体情况的检测结果. 可以看出相比于基于大数据的 K-means 的检测方法和基于大数据的 DBScan 检测方, 本文使用基于 Spark 的卷积神经网络的方法很大程度上提高了检测率和未知攻击检测率, 降低了误报率, 取得了比较好的成果.

表 4 各算法测试结果对比

类型	检测率			误报率			未知攻击检测率		
	K-means	DBScan	本文检测方法	K-means	DBScan	本文检测方法	K-means	DBScan	本文检测方法
DoS	82.08	87.01	98.04	17.92	12.99	3.15	72.54	83.69	94.74
R2L	83.44	88.45	95.41	16.56	11.55	4.04	71.97	82.21	95.18
U2R	85.07	89.17	96.59	14.93	9.83	3.26	70.73	80.64	94.85
综合	86.49	90.67	98.83	13.47	11.89	3.39	82.26	84.79	96.71

4 结论

本文结合电力控制网络的数据特点, 以及满足对精确性、实时性及效率的要求, 构建出了一个基于 Spark 的电网工业控制网络安全预警平台, 通过深度检测技术采集电力工控系统中的数据, 进行协议解析并存入 Kafka 相应话题, 然后利用 Spark Streaming 消费 Kafka 数据进行数据预处理并利用 Spark ML 与 Deep Learning4J 构建异常检测模型, 该方法在面对海量高维数据时不仅能快速有效地识别出异常流量的已有攻击以及未知攻击, 而且能够应用到各种工控系统, 体现了该系统的泛化能力与可扩展性高的优势. 本文方法能够做到实时处理及安全预警, 方便工作人员做出对应措施, 节约了人力成本, 极大地提高了工作效率, 提高了网络的安全性能.

参考文献

1 乔媛媛. 基于 Hadoop 的网络流量分析系统的研究与应用 [博士学位论文]. 北京: 北京邮电大学, 2014.

2 蒲晓川. 大数据环境下的网络流量异常检测研究. 现代电子技术, 2018, 41(3): 84–87.
3 吴晓平, 周舟, 李洪成. Spark 框架下基于无指导学习环境的网络流量异常检测研究与实现. 信息网络安全, 2016, (6): 1–7. [doi: 10.3969/j.issn.1671-1122.2016.06.001]
4 左晓军, 董立勉, 曲武. 基于 Spark 框架的分布式入侵检测方法. 计算机工程与设计, 2015, 36(7): 1720–1726.
5 王海凤. 工业控制网络的异常检测与防御资源分配研究[硕士学位论文]. 杭州: 浙江大学, 2014.
6 钟志琛. 基于网络流量异常检测的电网工控系统安全监测技术. 电力信息与通信技术, 2017, 15(1): 98–102.
7 Zaharia M, Chowdhury M, Franklin MJ, *et al*. Spark: Cluster computing with working sets. Proceedings of the 2nd USENIX Conference on Hot Topics in Cloud Computing. Boston, MA, USA. 2010. 10.
8 王震, 陈亮. 基于 Kafka 消息队列的电网设备准实时数据接入方法研究. 山东电力技术, 2015, 42(6): 41–43. [doi: 10.3969/j.issn.1007-9904.2015.06.009]
9 杨晨光, 马永征. 基于 Spark 的大规模网络流量准实时分类方法. 科研信息化技术与应用, 2016, 7(2): 25–34.
10 金振成. 基于 Spark Streaming 的 DDoS 攻击检测系统的设计与实现[硕士学位论文]. 北京: 北京交通大学, 2017.
11 Bell J. Machine Learning: Hands-on for Developers and Technical Professionals. Indianapolis: John Wiley & Sons, Inc., 2015. 275–314.
12 王伟. 基于深度学习的网络流量分类及异常检测方法研究[博士学位论文]. 合肥: 中国科学技术大学, 2018.
13 刘雨辰. 基于深度学习的路由器入侵检测技术研究[硕士学位论文]. 郑州: 战略支援部队信息工程大学, 2018.
14 赵彦辉, 范欣宁, 张建逵, 谢明. 基于 DeepLearning4J on Spark 深度学习方法在药用植物图像识别中应用初探. 中国中医药图书情报杂志, 2018, 42(5): 18–22. [doi: 10.3969/j.issn.2095-5707.2018.05.005]