

引用格式:李超鹏,王劲林. 基于两级分段模型的异构数据处理与网络攻击检测 [J]. 微电子学与计算机, 2019, 36(8): 54-58.
[LI C P, WANG J L. Heterogeneous data processing and network attack detection based on two-level and multi-segment model [J]. Microelectronics & Computer, 2019, 36(8): 54-58.]

基于两级分段模型的异构数据处理与网络攻击检测

李超鹏^{1,2}, 王劲林¹

(1 中国科学院声学研究所, 北京, 100190; 2 中国科学院大学, 北京, 100049)

摘 要: 由于网络数据存在着异构特性, 在网络攻击检测中, 需要使用合理的方法对异构数据进行整合. 我们在文章中提出了两级分段模型对异构数据进行整合并建模分析, 同时讨论了模型在多核条件下的模型分布式并行训练. 在实验部分, 本文介绍了使用的 DARPA 1998 网络攻击数据集, 并分析了不同分段方式下的攻击检测性能. 同时, 本文介绍说明模型在多核条件下的并行性能. 结果显示, 在面向异构数据时, 两级分段模型相比于无分段的全连接神经网络模型有着更好的检测准确率和召回率.

关键词: 攻击检测; 两级分段模型; 异构数据

中图分类号: TP393

文献标识码: A

文章编号: 1000-7180(2019)08-0054-05

Heterogeneous data processing and network attack detection based on two-level and multi-segment model

LI Chao-peng^{1,2}, WANG Jin-lin¹

(1 Institute of Acoustics Chinese Academy of Sciences, Beijing 100190, China; 2 University of Chinese Academy of Sciences, Beijing 100049, China)

Abstract: Because of the heterogeneous character of network data, a reasonable method is required to integrate the heterogeneous data in network attack detection. In this paper, we propose a two-level and multi-segment model for network attack detection. Meanwhile, the parallel training method is also involved. In the experiments, DARPA 1998 dataset has been used which is a public cyber attacks dataset. The experimental result shows that when facing the heterogeneous data, the two-level and multi-segments model has an effective performance that the proposed model obtains a better detecting accuracy and recall than the fully connected neural network.

Key words:

1 引言

在一般使用机器学习或深度学习模型检测网络攻击的研究中, 模型输入的网络攻击数据往往是报文数据^[1]. 报文数据作为输入, 一般有两种方式: 一种是通过网包协议解析, 提取特征字段, 输入模型进行训练. 这种方法可以看作是针对网络数据的特征工程^[2]. 在 SVM^[3]、聚类分析^[4] 等机器学习算法进

行攻击检测的研究中, 这种方法被大量采用; 另一种方法是将原始数据直接输入模型, 对于 IP 包或 TCP/UDP 包即以字节级别输入, 而不进行协议解析. 在许多深度学习模型进行网络攻击检测的工作中, 都以原始数据作为模型输入. Wang 等人使用 LSTM-CNN 模型^[5], 以原始数据作为输入进行攻击检测; Li 等人使用原始数据作为 RNN-RBM 模型的输入进行网络攻击检测^[6]; Jung 等人也使用了原

始数据作为 RNN 模型的输入^[7].

上述方法都是针对单一数据类型的检测模型,并没有解决网络异构数据问题^[8].为了解决这一问题,使得模型更有效的同时使用网络日志数据、报文数据和流量数据,基于网络安全态势感知的信息融合技术^[9]被提出.信息融合技术旨在将网络中的多种来源且异质的数据相融合,使检测系统能够综合各种数据源的信息,做出判断.信息融合技术的具体实现方法有许多,如 Bass 在 2000 年首次提出了在网络安全检测中应用多感知器的信息融合技术^[10];Giacobe 等人提出的基于 JDL 模型的信息融合攻击检测^[11];韦勇等人提出的将信息融合应用于安全态势评估^[12]等.上述方法大多数都是使用特征工程对异质的数据进行特征提取,之后输入模型进行分析.然而,特征工程的方法存在两个缺陷,一个是需要专家知识,另一个是特征提取的方法缺乏客观性,容易因人为偏差导致信息损失.

因此,本文提出了以原始数据作为输入的两级分段模型进行信息融合,从而解决网络攻击检测模型的异构数据问题.

2 面向异构数据的两级分段模型

2.1 网络数据的异构性

数据的异构性是指数据的类型不同,互相完全异质的特性.网络攻击检测主要数据源包括网络日志数据、网络报文数据和网络流量数据,其数据类型互不相同.我们将这种数据类型不同的现象称为异构数据.而单一检测模型无法同时处理多种数据类型的问题我们称为异构数据问题.在网络中,数据的异构性主要来源于网络设备差异.

2.2 两级分段模型

图 1 为两级分段模型的示意图.两级分段模型包含上级和下级.上级成为中心端神经网络,一般采用全连接的 MLP 模型.下级又成为边缘端神经网络.边缘端可以划分为多个分段,分段之间相互独立,每个分段都是一个小型的神经网络.

为了方便表示,此处假设边缘端模型为一个相同的两层 MLP 模型.假设边缘端划分为 M 个段,则每个海端的模型可以表示为 $f_m(\cdot)$, $m = 1, 2, \dots, M$.因此,边缘端模型表示为:

$$f_m(x) = \theta(W_{e2} \cdot \theta(W_{e1} x_m + b_{e1}) + b_{e2}) \quad (1)$$

式中, $\theta(\cdot)$ 是激活函数; W_{e1} 和 b_{e1} 是第一层神经网络的权重和偏置; W_{e2} 和 b_{e2} 是第二层神经网络的权重和偏置; x 是模型输入的向量.

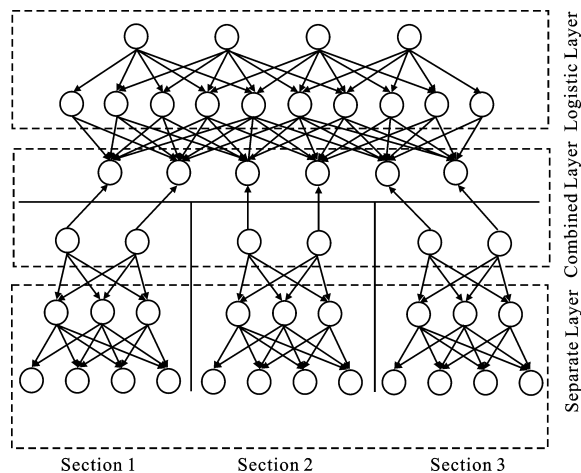


图 1 两级分段模型示意图

边缘端的 M 个输出向量将重新合并为一个新的向量,以 O_{edge} 表示一个新的向量:

$$O_{edge} = [f_1(x), f_2(x), \dots, f_M(x)] \quad (2)$$

那么,向量 O_{edge} 的长度为所谓子向量的和.以 O_{edge} 作为中心段的输入,可以得到中心段的神经网络的输出.本文假设中心段模型为一个 3 层全连接神经网络,则中心段的输出 O_c 为:

$$O_c = W_{c3} \theta(W_{c2} \theta(W_{c1} O_{edge} + b_{c1}) + b_{c2}) + b_{c3} \quad (3)$$

式中, W_{ci} 和 b_{ci} 是中心段第 i 层的权重和偏置; $\theta(\cdot)$ 是激活函数,对于中心端的输出向量,可以通过 Softmax 函数^[13]进行判别和分类.

2.3 两级分段模型的并行化

两级分段模型的各个边缘段子模型是相互独立的,中心端是依赖边缘端的输出.因此,两级分段模型可以进行模型分布式训练.本文基于多核 CPU 进行并行化训练分析.具体方案如图 2 所示.每个分段子模型由一个核负责,中心端的全局模型也由一个核负责,假设模型的分段数量为 M ,那么总共有 $M+1$ 个核心负责计算.

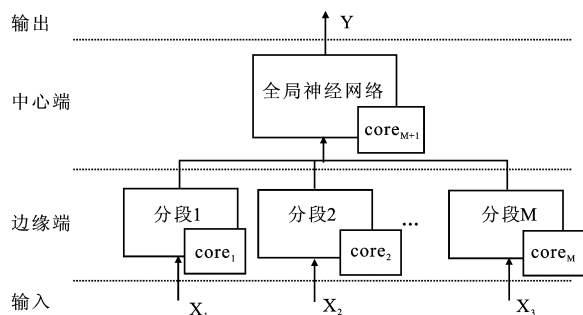


图 2 模型分布式的多核划分

两级分段模型的本质是一种模型分布式方法. 即每个计算核心负责一个子模型的计算, 子模型间的信息传递由核间通信实现.

3 实验分析

3.1 实验设置

3.1.1 实验数据

本文主要讨论网络报文数据和各种网络设备日志进行信息融合的问题. 本文采用 DARPA 1998 数据集^[14]进行建模分析. DAPPAR 1998 数据集是由林肯实验室于 1998 年在 DARPA 资助下开展了一项入侵检测评估项目, 项目的成果之一是建立了一个模拟各类入侵行为的流量数据集. 该数据集包含了 7 周的训练数据和 2 周的测试数据. 其中, 流量中包含了 4 种攻击类型, Dos, Probe, U2R, R2L. DAPPAR1998 的训练集中 attack 比例为 65.54%, 其中攻击类型占比分别为: 63.29%, 1.99%, 0.26%, 0.01%. 测试集数据比例与训练集相似.

对于网路报文数据, 可以分为 3 大类异构数据:

离散的报文数据. 我们以 IP 包报文作为处理对象, 以字节为单位进行处理并作为两级分段模型的输入. IP 报文头解析后, 可以得到若干特征字段, 都为离散值. 在作为模型输入时, 一般重整为浮点型数据, 但由于离散数据特征较为明显, 一般采用层数较浅的神经网络模型进行处理. 因此, 在处理报文数据的边缘端神经网络层数较少.

文本形式的日志数据. 本文通过 Snort^[15] 开源攻击检测软件, 对 DARPA 1998 的数据进行回放, 并使用 Snort 软件进行检测, 得到对应的日志数据. 对于 Snort 的日志数据, 大多数是以文本的形式存在. 这类数据无法作为神经网络的输入. 因此, 一般采用词嵌入的方法进行处理. 本文采用 word2vec 工具^[16]对所有日志中的单词计算编码值. Word2vec 是 Google 开发的一款开源词向量工具, word2vec 利用语料库数据进行训练, 得到大多数常见单词的词向量. 本文采用 300 维长度的词向量. 由于日志的单词数量不是固定的, 而一则日志记录为一句或若干句组成. 因此本文采用所有词向量平均的方法构造一则日志记录的向量.

连续的流量数据. 另外一类数据为网络流量数据. 通过对 DARPA 1998 的原始数据报文进行统计分析, 可以得到 7 周的网络流量变化. 这类数据是以连续值形式存在的.

3.1.2 评估指标

本文中, 我们主要采用 2 种评估指标: 准确率和召回率. 准确率是描述系统误差的指标, 常以分类器的正确率作为表示. 然而, 如果一些攻击非常重要, 那么召回率, 一种表示检测到的样例占总体样例的比值, 将被关注. 这 2 种指标表示如下:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \quad (4)$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (5)$$

式中, TN 是正确被预测为非攻击的样本数量; FN 是错误被预测为非攻击的样本数量; FP 是错误被预测为攻击的样本数量; TP 是正确被预测为攻击的样本数量. 在网络攻击检测中, 我们更加关注攻击行为, 因此恶意样本被视为正样本. 值得一提的是, 上述所有评估指标都是基于测试集计算的.

3.2 异构数据分段的影响

根据数据的异构性, 我们将考虑将异质数据分别通过不同的边缘端子模型进行处理, 得到特征向量后重新输入中心端模型. 然而, 为了说明异构数据分离处理的有效性, 我们设计了不同的分段方法.

对照组: 模型边缘端仅 1 个分段, 即不进行分段. 所有数据在边缘端模型处理后直接输入中心端.

实验组 1: 模型边缘端共 2 个分段, 第 1 个分段以报文数据作为输入, 第 2 分段以日志数据和流量数据作为输入.

实验组 2: 模型边缘端共 2 个分段, 第 1 个分段以日志数据作为输入, 第 2 分段以报文数据和流量数据作为输入.

实验组 3: 模型边缘端共 2 个分段, 第 1 个分段以流量数据作为输入, 第 2 分段以报文数据和日志数据作为输入.

实验组 4: 模型边缘端共 3 个分段, 第 1 个分段以报文数据作为输入, 第 2 分段以报文数据作为输入, 第 3 分段以日志数据作为输入. 具体分段方法见表 1. 该设计方法目的是通过合并与分离不同的数据类型, 观察模型最终输出的攻击检测准确率和召回率, 确定分段的影响. 需要注意的是, 对照组本质上是多层全连接神经网络. 最终结果见图 3.

对照组准确率为 81.65%, 召回率为 77.59%. 而实验组 4, 即 3 个分段的两级分段模型, 准确率为 98.30%, 召回率为 98.08%, 分别提升了 16.65% 和 20.41%. 可以看出, 实验组 4 的各项指标都明显优于其它实验组和对照组. 而在实验组 1 到实验组 3 中, 我

们发现,实验组 2 的效果略优于实验组 1 和 3,说明日志数据的异质化程度要强于报文数据和流量数据.

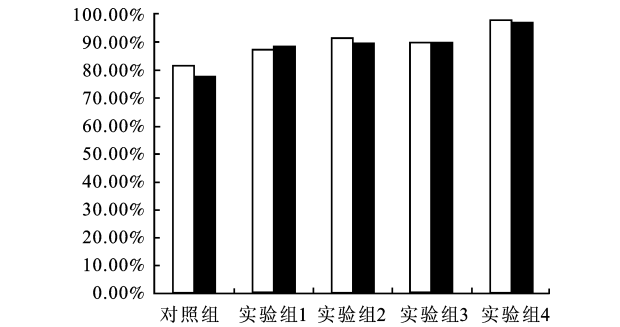


图 3 实验组准确率及召回率

两级分段模型的提出,主要目的是解决异构数据处理的问题. 在使用普通全连接深度神经网络进行异构数据建模的情况下,模型的准确率较低,而通过两级分段模型将异构数据分别进行建模再融合后,得到的准确率有所提高. 但是,为了进一步验证模型的有效性,我们考虑在 3 个分段的基础上继续

进行分段实验,确认更多的分段是否会继续提高模型的准确率. 实验结果如图 4 所示.

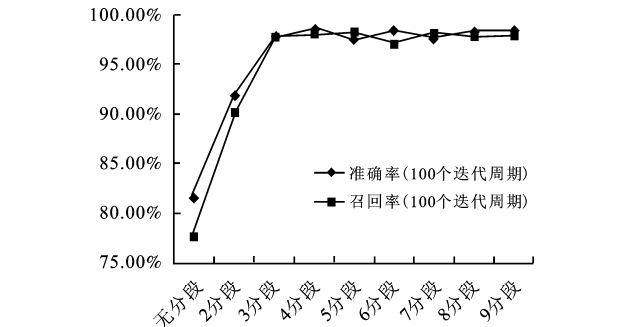


图 4 1 到 9 分段下的模型准确率和召回率对比

可以看出,更多的模型分段并没有提高模型的准确率,相对的甚至可能因为过多的分段导致模型的权重数量过少,模型连接过于稀疏,导致模型在训练过程中过快收敛于局部最优解,最终使得模型准确率或召回率的下降.

表 1 对照组及实验组设置

实验组名称	分段数量	数据划分			神经元数量	
		第 1 段	第 2 段	第 3 段	边缘端第 1 层	边缘端第 2 层
对照组	1	全数据	——	——	500	300
实验组 1	2	报文	日志+流量	——	100+400	60+240
实验组 2	2	日志	报文+流量	——	300+200	180+120
实验组 3	2	流量	报文+日志	——	100+400	60+240
实验组 4	3	报文	日志	流量	100+300+100	60+180+60

注:中心端每层神经网络规模为 300/300/100,由于各组一致,不在表中列出.

3.3 两级分段模型的并行性能

基于前文所述的两级分段并行化算法,可以对模型进行分布式运行. 通过实验,可以得到模型并行化的运行时间. 为了确认并行化的效率,本文提出在同一数据类型内部进一步做分段处理,使得分段数量超过 3 段. 故并行性能实验共测试 1 到 10 段分段情况的运行时间. 结果如图 5 所示.

由于本文采用的 8 核的 CPU,当超过 8 个分段数之后,多余的分段之间只能由同一个核心处理. 可以看出,在完全使用多核 CPU 进行模型训练下,模型的并行训练时间相比于单核训练时间,有着明显的下降. 当分段过超过 5 之后,模型的运行效率没有明显提升. 随着分段的增加,模型并行的性能也不是成比例加速的,主要原因是模型训练过程中,存在大量数据通信的开销,因此过多的分段可能增加了通信开销和核间通信时间.

另外,单核训练的时间随着分段数的增加,时间

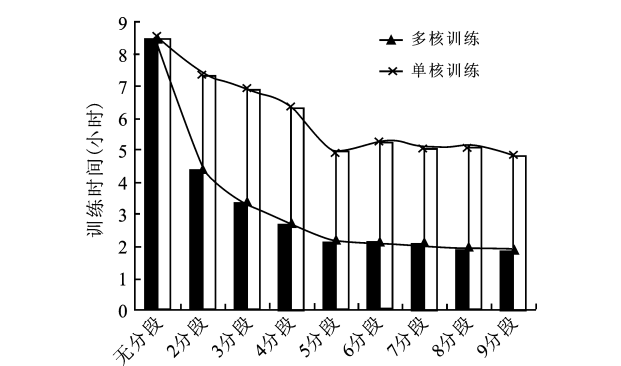


图 5 多分段下的训练时间

也有所减少. 主要原因是,在模型的神经元数量不变的前提下,分段增加,将减少权重的数量,使得神经网络的连接更加稀疏,计算量下降,故训练时间减少.

4 结束语

本文旨在提出一种解决网络数据异构特性的攻

击检测模型. 通过本文提出的两级分段模型, 解决了报文、日志和流量数据的异构特性, 通过边缘端分别处理异质数据, 得到不同数据的特征后, 再通过中心端模型统一处理进行攻击检测. 实验证明, 本文提出的模型有着较好的检测效果, 并且在多核条件下的模型分布式运算能够有效提高训练效率. 当然, 该模型仍然有着改进的空间. 在未来的工作中, 我们计划研究不同分段使用不同层数的模型, 确定异构数据对应的最优神经网络层数; 另外, 我们计划分析报文、日志和流量数据, 确定异质的网络数据对攻击检测结果的影响程度.

参考文献:

- [1] Stallings W. Business data communications[M]. Prentice Hall PTR, 1990.
- [2] Ng A. Machine Learning and AI via Brain simulations[J]. 2013.
- [3] Xu X. Adaptive intrusion detection based on machine learning: feature extraction, classifier construction and sequential pattern prediction[J]. International Journal of Web Services Practices, 2006, 2(1-2): 49-58.
- [4] 马晓春, 高翔, 高德远. 聚类分析在入侵检测系统中的应用研究[J]. 微电子学与计算机, 2005, 22(4): 134-136.
- [5] Wang W, Sheng Y, Wang J, et al. HAST-IDS: learning hierarchical spatial-temporal features using deep neural networks to improve intrusion detection[J]. IEEE Access, 2018, 6: 1792-1806.
- [6] Li C, Wang J, Ye X. Using a Recurrent Neural Network and Restricted Boltzmann Machines for Malicious Traffic Detection[J]. NeuroQuantology, 2018, 16(5).
- [7] Jung W, Kim S, Choi S. Poster: deep learning for zero-day flash malware detection[C]//36th IEEE symposium on security and privacy. 2015.
- [8] 张宇, 蒋东兴, 刘启新. 基于元数据的异构数据集整合方案[J]. 清华大学学报: 自然科学版, 2009, 49(7): 1037-1040.
- [9] Blasch E, Bossé ?, Lambert D A. High-level information fusion management and systems design[M]. Artech House, 2012.
- [10] Bass T. Intrusion detection systems and multisensor data fusion[J]. Communications of the ACM, 2000, 43(4): 99-105.
- [11] Giacobe N A. Application of the JDL data fusion process model for cyber security[C]//Multisensor, Multisource Information Fusion: Architectures, Algorithms, and Applications 2010. International Society for Optics and Photonics, 2010, 7710: 77100R.
- [12] 韦勇, 连一峰, 冯登国. 基于信息融合的网络安全态势评估模型[J]. 计算机研究与发展, 2009, 46(3): 353-362.
- [13] Nasrabadi N M. Pattern recognition and machine learning[J]. Journal of electronic imaging, 2007, 16(4): 049901.
- [14] Lippmann R, Cunningham R K, Fried D J, et al. Results of the DARPA 1998 Offline Intrusion Detection Evaluation[C]//Recent advances in intrusion detection. 1999, 99: 829-835.
- [15] Roesch M. Snort: Lightweight intrusion detection for networks[C]//Lisa. 1999, 99(1): 229-238.
- [16] Mikolov T, Chen K, Corrado G, et al. Efficient estimation of word representations in vector space[J]. arXiv preprint arXiv:1301.3781, 2013.

作者简介:

李超鹏 男, (1990-), 博士研究生. 研究方向为网络安全和深度学习. E-mail: licp@dsp.ac.cn.

王劲林 男, (1964-), 研究员. 研究方向为数字信号处理、信源编码和信道编码、网络中基于媒体流的应用技术以及 5G 通信.