doi:10.11835/j.issn.1000-582X.2021.11.010

基于 Transformer 神经网络模型的网络入侵检测方法

郭志民[1]，周劼英[2]，王　丹[3]，吕　卓[1]，杨　文[1]

（1.国网河南省电力公司电力科学研究院，郑州 450000；2.国家电网有限公司，北京 100031；
3.国网河南省电力公司，郑州 450000）

摘要：网络入侵检测一直以来都是网络安全中亟待解决的关键任务之一，传统网络入侵检测方法主要通过提取多维特征，采用机器学习方法构建检测模型，大多忽略了入侵行为的时间相关性。通过提取网络入侵行为的时序特征，设计基于降维特征的多头自注意力机制 Transformer 网络模型，以解决传统串行化时序神经网络模型不易收敛且时间开销较大的问题，通过选取最优的损失函数和训练参数进行并行化训练，实现网络入侵行为检测。实验结果表明，基于 Transformer 网络模型的网络入侵检测方法在多个数据集上均获得了 99%以上的精度和检出率。

关键词：入侵检测；时序神经网络；Transformer；自注意力机制；深度学习

中图分类号：TP391　　**文献标志码**：A　　**文章编号**：1000-582X(2021)11-081-08

Network intrusion detection method based on Transformer neural network model

GUO Zhimin[1]，*ZHOU Jieying*[2]，*WANG Dan*[3]，*LV Zhuo*[1]，*YANG Wen*[1]

(1. State Grid Henan Electric Power Research Institute, Zhengzhou 450000, P. R. China; 2. State Grid, Beijing 100031, P. R. China; 3. State Grid Henan Electric Power Company, Zhengzhou 450000, P. R. China)

Abstract: Network intrusion detection has always been one of the key tasks in network security. Traditional network intrusion detection methods mainly use machine learning method to construct detection models by extracting multi-dimensional features, while most of them ignore the time correlation of intrusion behaviors. In this paper, a Transformer network model with multi-head self-attention mechanism based on dimension reduction feature was designed by extracting the time sequence features of network intrusion behavior. The proposed model solved the problems that traditional serial sequential neural network models are difficult to converge and have a large time consumption. The optimal loss function and training parameters were selected to implement the network intrusion detection. The experimental results show that the network intrusion detection method based on Transformer network model achieves the accuracy and the detection rate of over 99% in multiple datasets.

Keywords: intrusion detection; recurrent neural network; Transformer; self-attention mechanism; deep learning

收稿日期：2021-05-18

基金项目：国家电网有限公司总部科技项目资助(5700-202024193A-0-0-00)。

Supported by the Science and Technology Project of State Gird Gorporation of China(5700-202024193A-0-0-00).

作者简介：郭志民(1977—)，男，教授级高级工程师，主要从事信息安全方向研究，(E-mail)zhimin.guo@163.com。

当今网络安全形势日益严峻,网络攻击者利用巧妙的攻击手法避开防火墙,入侵网络系统获取隐私信息、破坏网络系统或导致服务器瘫痪,关于网络入侵检测研究已成为当今网络安全最重要的研究方向之一。网络入侵检测通过对计算机系统和网络事件分析,检测入侵和攻击行为。在一个网络系统中,任何未经授权的活动,以及企图绕过网络安全机制的行为,都可视为网络入侵行为[1]。网络入侵检测可被分为基于异常的检测和基于误用的检测2种[2-3]。基于异常的检测系统通过观察网络、系统或用户的异常行为来检测攻击行为,基于误用的检测系统则使用先验的攻击模式和签名来检测攻击行为。

随着人工智能和机器学习的发展,越来越多研究开始尝试使用机器学习的方法解决网络入侵检测的难题。Chowdhur等人[4]以互联网上的流量数据为训练集,一次性选取任意3组特征作为SVM的输入进行训练,给予了SVM一定检测任意网络异常行为的能力。Mohsen等人提出了用于入侵检测的最小-最大K均值聚类方法[5]。该算法试图最小化簇的最大内部方差,而不是像K均值算法那样最小化内部方差的和。每个集群都有一定的权重,并将较高的权重分配给内部方差较大的集群,该算法获得了81%的检测率。Li等人[6]提出了2阶段的"智能入侵检测方法"。第一阶段包括使用随机森林算法,通过权衡特征的重要性来获得特征的子集。第二个阶段是一种基于特征子集作为输入的分类器"基于混合聚类的Adaboost算法"。Jaiganesh等人[7]提出一种基于神经网络的入侵检测算法,通过专门选取入侵数据,使用反向传播算法训练神经网络权重,使算法具有检测入侵行为的能力。Sinapiromsaran等人提出了多属性框架决策树[8],将数据分为左、中、右3个区域,从最远的一对中选择一个核心向量来对入侵行为进行分类。李俊等人[9]考虑了网络入侵数据的时序特点,使用GRU_RNN网络结构在KDD数据集上进行训练,获得比其他非时序网络更好的识别率与收敛性。

尽管许多现有研究探索了机器学习在网络入侵检测中的应用,这些研究对正常行为和攻击行为进行分类,基于机器学习方法构建入侵检测模型,具有一定检测效果,但仍存在一些问题。主要体现在:1)训练样本中标签为正常行为的数据量远大于非法行为,数据特征分布严重不均,导致模型难以训练且泛化能力不足。2)网络入侵通常是时间上的一段连续行为,大多数模型不具备时序学习能力而丢失了时序特征,部分基于循环神经网络的方法虽能够学习时序特征,但其基于序列的串行训练方式存在训练耗时长且收敛效率较低等问题。

Transformer[10]最初应用于自然语言处理(NLP)任务中,其结构完全抛弃了RNN和CNN等网络结构,而仅采用Attention机制来进行机器翻译任务,且取得了很好效果,其网络结构如图1所示。Devlin等人提出的BERT[11],Brown等人提出的GPT-3[12],这些基于Transformer的模型都在NLP领域取得了重大突破。Transformer与基于RNN的时序神经网络有明显不同,RNN的训练是迭代的、串行的,而Transformer的训练是并行的,即所有特征是同时训练的,大幅增加计算效率。

通过分析网络入侵行为的数据特征,提出基于Transformer神经网络模型的入侵检测方法。通过在多个数据集上进行实验,选取最优的损失函数和网络结构,最后在测试数据集上,相较于对比机器学习方法,提升训练效率和识别率。主要贡献包括:

1)针对网络入侵行为数据的时间相关性,提出了一种基于Transformer的网络入侵检测方法,进一步提升网络入侵检测的准确性。

2)设计一种基于降维特征的多头自注意力机制Transformer网络模型,以解决传统串行化时序神经网络模型不易收敛且时间开销较大问题,通过选取最优损失函数和训练参数进行并行化训练,从而实现网络入侵行为检测。

3)在多个数据集上进行对比实验,结果表明,提出的基于Transformer网络模型的网络入侵检测方法在多个数据集上均获得了99%以上的精度和检出率。

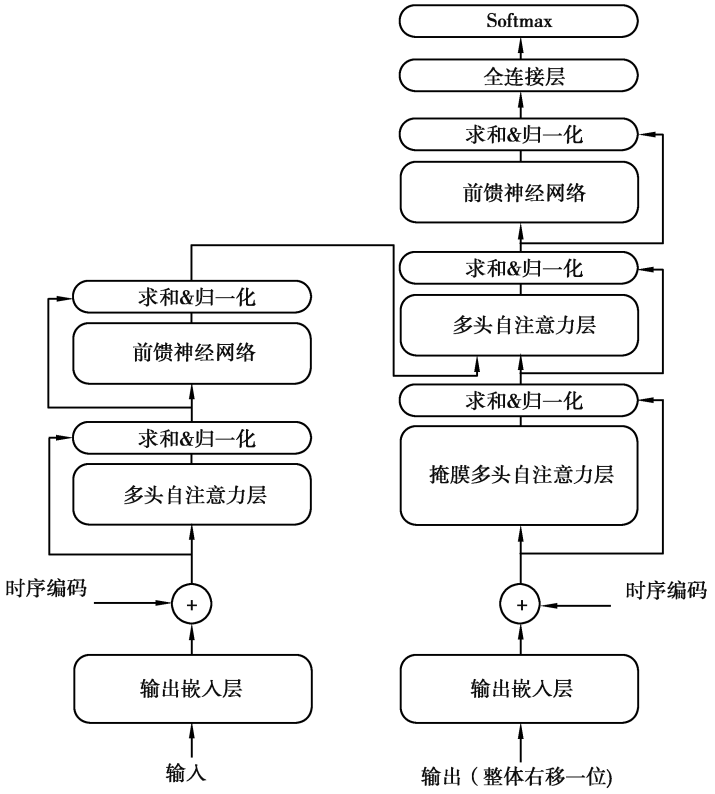

图 1　Transformer 网络结构

Fig. 1　Transformer network structure

1　数据分析及预处理

1.1　数据分析

实验采用的数据集为 KDD-Cup-99 和 NSL-KDD 网络入侵数据集。KDD-Cup-99 数据集[13]是第三届国际知识发现和数据挖掘工具竞赛所使用的数据集，共计 23 种标签、4898431 条数据，包含正常和 22 种攻击类型标签。NSL-KDD 数据集[14]是 KDD-Cup-99 数据集的改进版本，包含 125973 条网络连接记录。数据集如表 1 所示。

表 1　网络入侵数据集

Table 1　Network intrusion data set

数据集名称	数量
KDD-Cup-99	4 898 431
NSL-KDD	125 973

分析了 KDD-Cup-99 及 NSL-KDD 数据集的数据分布，分析结果如图 2 所示。

分析结果表明，KDD-Cup-99 数据集分布不平衡，这种不平衡数据分布会导致模型性能欠佳，导致漏检率升高，而 NSL-KDD 数据存在信息冗余的问题[15]。为解决这一问题，引入特征提取模型，通过数据表征和降维，避免因数据冗余造成收敛性能降低。

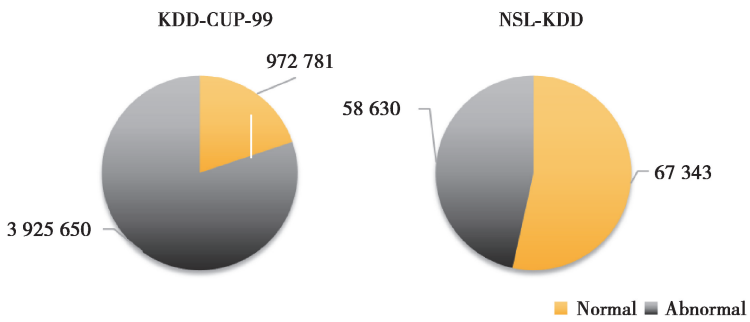

图2 KDD-Cup-99及NSL-KDD数据分布

Fig. 2 Data distribution of KDD-Cup-99 and NSL-KDD

1.2 数据特征提取

1.2.1 字符数据编码

由于原始数据集包含了字符串特征,不利于直接向量化,为了方便计算,将数据标签进行One-hot编码。One-hot编码是机器学习分类任务中常用的数据编码方式,它可以将原数据中离散的值转化为欧式空间的点,使各标签之间保持合理的特征距离。数据集中的每条数据被分为正常或异常2种类别,正常编码为01,异常编码为10,具体编码方式见表2所示。

表2 数据标签编码

Table 2 Data label code

标签	编码
Normal	01
Abnormal	10

1.2.2 归一化

由于原始数据中数据范围相差较大,不利于网络训练。所以需要对原始数据的每一列进行归一化处理。将同一列数据归一化到(0,1)之间。其归一化公式为

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

其中:x 为原始数据集的任意一列数据值;$x_{\min}$ 为统计整列获得的最小值;$x_{\max}$ 为最大值;x^* 为归一化后的数据值。

1.2.3 特征降维

为了去除数据集中冗余信息对检测准确性的影响,引入一个特征提取网络 F 作为入侵检测模型的前置网络,该网络由2层全连接层构成,其目的是将冗余的低级特征映射为高级特征。特征提取网络 F 的计算过程如公式2所示。

$$y = \sigma(x') = \sigma(F(\boldsymbol{x}^*)), \tag{2}$$

其中:x^* 为获得的归一化数据;特征长度为 d_x^*;特征提取网络输出其高级特征向量 $\boldsymbol{x}'$;特征长度为 d_x,设定 $d_x < d_x^*$;σ 为激活函数。将输出层未激活的特征向量 $\boldsymbol{x}'$ 作为高级映射,通过在整个原始数据集上训练网络 F,即可得到新的特征数据集 D,其中特征数据集 D 是由特征长度为 d_x 的特征向量构成。

2 基于Transformer网络模型的网络入侵检测

2.1 时序编码

首先对网络入侵数据进行时序编码,对于特征数据集 D,需要将时序信息嵌入到输入特征中,通过一层全连接层,对不同时间的特征进行相应的时序编码。时序编码计算公式如下

$$PE_{(\mathrm{pos},2i)}=\sin(\mathrm{pos}/10\ 000^{2i/d_x}),\tag{3}$$

$$PE_{(\mathrm{pos},2i+1)}=\cos(\mathrm{pos}/10\ 000^{2i/d_x}),\tag{4}$$

其中：pos 指的是一段序列中某个时刻特征的位置，取值范围为[0，最大序列长度]；d_x 是特征维度；i 表示在时序编码向量中的索引，取值范围为$[0,\ldots,d_x]$。位置嵌入函数的周期从 2π 到 10 000×2π 变化，而每一个位置在编码维度上都会得到不同周期的 sin 和 cos 函数的取值组合，从而产生独一的纹理位置信息，最终使得模型学到位置之间的依赖关系和自然语言的时序特性[4]。

2.2　编解码模块

从时序编码后的特征序列中取一个长度为 t 的连续序列 $X_t=\boldsymbol{x}_1,\ldots,\boldsymbol{x}_t\mid x_i\in R^{d_x}$，其中 $\boldsymbol{x}_t$ 是在 t 时刻维度为 d_x 的网络信息特征向量。输出为一个长度为 t 的 $Y_t=y_1,\ldots,y_t\mid y_i\in(0,1)$状态集，其中 y_t 是在 t 时刻的状态。输入数据经过时序编码后进入编码模块，编码模块将特征映射到更高维的特征图（如图 3 所示），并将其输入到解码模块中，解码模块输出最终的状态集。

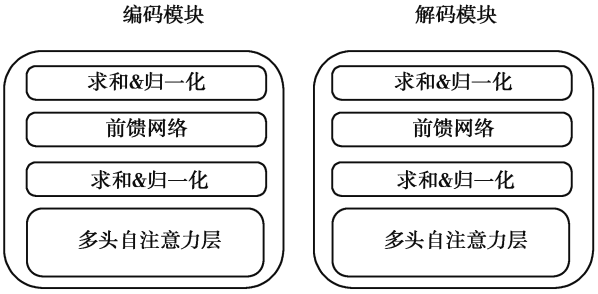

图 3　编码模块及解码模块结构图

Fig. 3　Structure of encoding module and decoding moudle

编码模块与解码模块具有相同的结构，主要由多头自注意力层和前馈网络组成。为了让模型去关注不同方面的信息，采用了多头自注意力层将注意力模块分为多个头，从而产生多个子空间，增强模型性能。

2.3　网络结构设计

网络由编码器和解码器组成（如图 4 所示），其中编码器主要由输入层，时序编码层和编码模块组成，输入层通过全连接层将输入时间序列数据映射到高维的向量，然后将输入的向量与时序编码向量逐元素相加，对其特征进行时序编码。然后将结果输入到编码层，在经过编码器后生成的特征向量，将其送入解码器中。

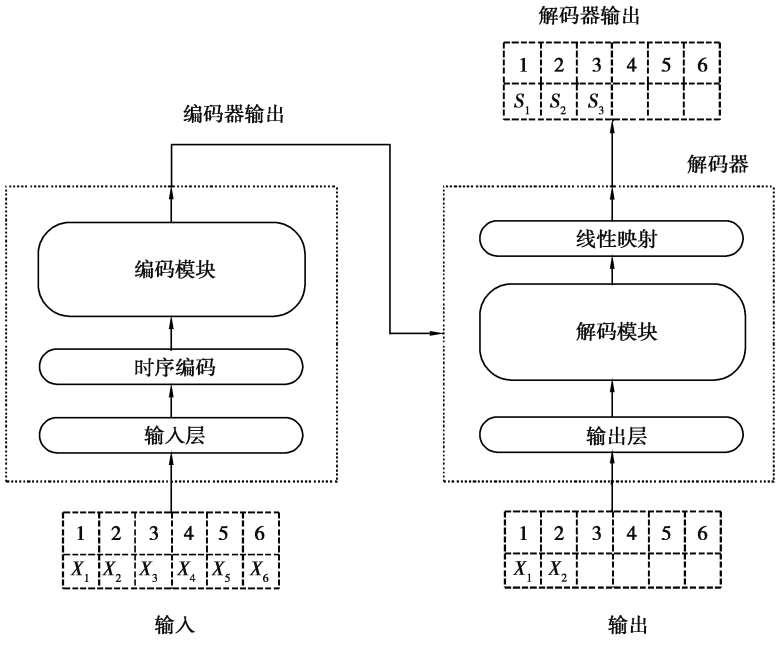

图 4　基于 Transformer 网络入侵检测模型

Fig. 4　Network intrusion detection model based on transformer

在推理的过程中采用动态解码的方式，在经过编码器之后获得高维特征图，将其输出到解码器中，同时根据前面时刻的预测结果依次进行。

2.4 损失函数

由于数据样本中存在正负样本不均衡的问题，采用 Focal Loss[16] 作为损失函数，如公式(5)所示，其广泛用于目标检测任务中的困难样本挖掘，通过调整正负样本的权重，使得模型在训练中更关注难分类的样本，有效缓解数据分布不均问题。

$$L_{fl}=\begin{cases}-\alpha(1-y)\gamma\log y, & y'=1\\ -(1-\alpha)y^{\gamma}\log(1-y), & y'=0\end{cases}, \tag{5}$$

其中：y 为分类层激活函数的输出；y'为真实值，即编码后的标签；α 和 γ 为调节因子，α 取值为 0.25，γ 取值为 2。

3 实验分析

3.1 实验环境

本次实验中采用的硬件环境配置为 Intel(R) Core(TM) i7-9700 CPU 64 位处理器、32 GB 内存，并采用 GTX 3080 运算加速，操作系统为 Ubuntu 16.04。按比例 4∶1随机拆分训练集和测试集，优化器采用 Adam，设置初始学习率为 0.001，epoch 数为 100。

3.2 验证方法

为了对实验结果进行有效性能评估，采用二分类任务评价的标准混淆矩阵，混淆矩阵如表 3 所示。

表 3 混淆矩阵

Table 3 Confusion matrix

检测结果 / 实际情况	检测为入侵	检测为正常
实际为入侵	TP	FP
实际为正常	FN	TN

根据混淆矩阵，可以得到以下 3 个评价指标包括精度(PRE)、检出率(TPR)、F1 分数，如公式(6)、(7)和(8)所示。

精度(PRE)

$$\mathrm{PRE}=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}}, \tag{6}$$

检出率(TPR)

$$\mathrm{TPR}=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}, \tag{7}$$

F1 分数

$$\mathrm{F1}=\frac{2\cdot\mathrm{PRE}\cdot\mathrm{TPR}}{\mathrm{PRE}+\mathrm{TPR}}。 \tag{8}$$

3.3 性能评估

与传统方法机器学习方法、基于深度学习的机器学习方法进行对比分析实验[17-21]。传统方法，与基于特征提取的支持向量机(SVM)算法及基于聚类的最邻近结点算法(KNN)算法进行对比。深度学习方法，与深度神经网络(DNN)和基于递归神经网络的长短时记忆神经网络(LSTM)进行对比。在不同数据集上，采用精度、检出率和 F1 分数 3 个准确性指标进行对比实验，验证相对于其他模型，基于 Transformer 网络模型的检测方法在检测效果上的优势。图 5、图 6 分别展示了与传统方法和与深度学习方法的对比实验结果。

1)与传统方法相比，提出的检测方法在精度和检出率方面都有明显优势，SVM 相较 KNN 提升一定检测

效果，但仍不佳。在不同数据集上，传统方法检测效果受影响较大，而研究方法无论是在数据分布不均的 KDD-Cup-99 数据集上，还是在数据相对分布均匀的 NSL-KDD 数据集上，都能取得好的检测效果。

2）与深度学习方法进行对比实验，采用相同的模型，使用不同的数据集训练，DNN 与 LSTM 的检测效果也会受到影响，各指标波动明显大于检测模型，这表明方法在模型泛化性能上更具优势。采用相同的训练集训练，使用不同的模型进行对比，深度学习模型均能取得 95％以上的精度和检出率，且具有时序学习能力的 LSTM 比 DNN 有更好的准确性，说明网络入侵行为存在可用的时序信息，提出检测方法在各指标上取得了 99％以上分数，优于其他深度学习模型。实验结果表明，针对数据分布、时序信息学习及网络结构改进都有效提升网络入侵检测效果。

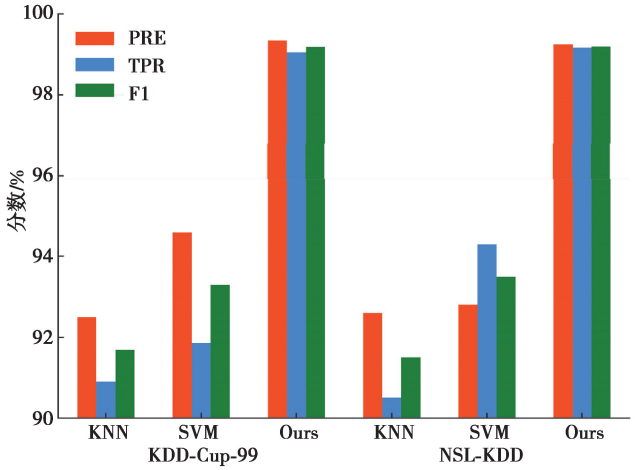

图 5　传统方法对比实验结果

Fig. 5　Comparison experimental results of traditional methods

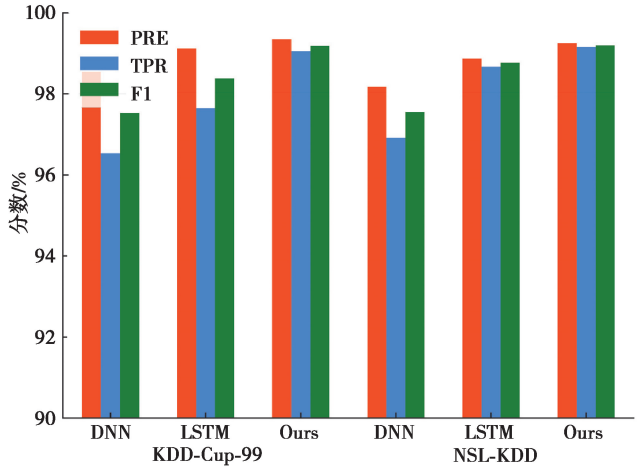

图 6　深度学习方法对比实验结果

Fig. 6　Comparison experimental results of deep learning methods

4　结　语

提出一种基于 Transformer 网络模型的网络入侵检测方法。所提出的 Transformer 网络模型基于降维特征，利用多头自注意力机制学习到网络入侵数据时序特征，通过选取最优的损失函数和训练参数进行并行化训练，结合特征提取的数据预处理方式，缓解数据分布不均衡问题，有效提高检测效果。实验结果表明，在不同数据集上，相比传统方法以及深度学习方法，采用精度、检出率和 F1 分数作为指标，都取得了最佳检测效果。

参考文献：

[1] Liao H J, Lin C, Lin Y C, et al. Intrusion detection system: A comprehensive review[J]. Journal of Network & Computer Applications, 2013, 36(1):16-24.

[2] Ibrahim L M. Artificial neural network for misuse detection[J]. Journal of Communication and Computer, 2010, 7(6): 38-48.

[3] Rai M, Mandoria H L. Network intrusion detection: a comparative study using state-of-the-art machine learning methods [C]// 2019 International Conference on Issues and Challenges in Intelligent Computing Techniques (ICICT), September 27-28, 2019, Ghaziabad, India: IEEE, 2019.

[4] Chowdhury M N, Ferens K, Ferens M. Network intrusion detection using machine learning[C]// Proceedings of the International Conference on Security and Management (SAM). The Steering Committee of The World Congress in Computer Science, Computer Engineering and Applied Computing (WorldComp), August 13-14, 2016, Guilin, China: Atlantis Press, 2016: 30.

[5] Mohsen E, Varjani A Y. Intrusion detection based on MinMax K-means clustering[C]// 7th International Symposium on

Telecommunications (IST'2014), September 9-11, 2014, Tehran, Iran: IEEE, 2014: 804-608.

[6] Li J, Zhao Z, Li R. A machine learning based intrusion detection system for software defined 5G network[J]. arXiv preprint arXiv:1708.04571, 2017.

[7] Jaiganesh V, Sumathi P, Mangayarkarasi S. An analysis of intrusion detection system using back propagation neural network[C]// 2013 international conference on information communication and embedded systems (ICICES), February 21-22, 2013, Chennai, India: IEEE, 2013:232-236.

[8] Sinapiromsaran K, Techaval N. Network intrusion detection using multi-attributed frame decision tree[C]// 2012 Second International Conference on Digital Information and Communication Technology and it's Applications (DICTAP), May 16-18, 2012, Bangkok, Thailand: IEEE, 2012:203-207.

[9] 李俊, 夏松竹, 兰海燕, 等. 基于 GRU-RNN 的网络入侵检测方法[J]. 哈尔滨工程大学学报, 2021, 42(6): 879-884.
LI J, XIA S Z, LAN H Y, et al. Network intrusion detection method based on GRU-RNN[J]. Journal of Harbin Engineering University, 2021, 42(6): 879-884.(in Chinese)

[10] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. arXiv 2017[J]. arXiv preprint arXiv:1706.03762, 2017: 2999-3007.

[11] Devlin J, Chang M W, Lee K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding [J]. arXiv preprint arXiv:1810.04805, 2018.

[12] Brown T B, Mann B, Ryder N, et al. Language models are few-shot learners[J]. arXiv preprint arXiv: 2005.14165, 2020.

[13] KDD Cup 1999[DB/OL]. (1999-10-28)[2021-05-06]. http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.

[14] The NSL KDD Dataset[DB/OL]. (2013-7-30)[2021-05-06].http://nsl.cs.unb.ca/NSL-KDD/.

[15] Zhang X Y, Zeng H, Jia L. Research of intrusion detection system dataset-KDD CUP99[J]. Computer engineering and design, 2010, 31(22): 4809-4812.

[16] Lin T Y, Goyal P, Girshick R, et al. Focal loss for dense object detection[C]// Proceedings of the IEEE international conference on computer vision, October 22 - 29, 2017, Venice, Italy: IEEE, 2017: 2999-3007.

[17] Alazzam H, Sharieh A, Sabri K E. A feature selection algorithm for intrusion detection system based on pigeon inspired optimizer[J]. Expert Systems with Applications, 2020, 148: 113249.

[18] Xie M, Hu J, Yu X, et al. Evaluating host-based anomaly detection systems: application of the frequency-based algorithms to ADFA-LD[C]// International Conference on Network and System Security, November 3-5, 2015, New York, USA: Springer, Cham, 2015: 542-549.

[19] 侯湘, 黄晋, 桑军, 夏晓峰. 多维度融合的文献作者亲密度计算[J]. 情报学报, 2021, 40(8): 846-853.
H X, H J, S J, X XF. Calculation of author intimacy based on multi-dimensional fusion. 情报学报, 2021, 40(8): 846-853.(in Chinese)

[20] Javaid A, Niyaz Q, Sun W, et al. A deep learning approach for network intrusion detection system[C]// Proceedings of the 9th EAI International Conference on Bio-inspired Information and Communications Technologies (formerly BIONETICS), December 3-5, 2015, New York, United States: EAI, 2016: 21-26.

[21] Althubiti S, Nick W, Mason J, et al. Applying long short-term memory recurrent neural network for intrusion detection[C]// SoutheastCon 2018, April 19-22, 2018, Florida, USA: IEEE, 2018: 1-5.

(编辑 侯 湘)