基于分层自编码器的异常网络流量检测

Malicious Network Traffic Detection Based on Hierarchical Autoencoder

张晓青/ZHANG Xiaoqing[1,2]，谷勇浩/GU Yonghao[1,2,3]，田甜/TIAN Tian[4]
（1. 北京邮电大学计算机学院，中国 北京 100876；
2. 北京邮电大学智能通信软件与多媒体北京市重点实验室，中国 北京 100876；
3. 中山大学广东省信息安全技术重点实验室，中国 广州 510275；
4. 中兴通讯股份有限公司，中国 深圳 518057）
(1. School of Computer Science, Beijing University of Posts and Telecommunications, Beijing 100876, China;
2. Beijing Key Laboratory of Intelligent Telecommunications Software and Multimedia, Beijing University of Posts and Telecommunications, Beijing 100876, China;
3. Guangdong Provincial Key Laboratory of Information Security Technology, Sun Yat-Sen University, Guangzhou 510275, China;
4. ZTE Corporation, Shenzhen 518057, China)

DOI：10.12142/ZTETJ.202405012
网络出版地址：http://kns.cnki.net/kcms/detail/34.1228.TN.20220706.0920.002.html
网络出版日期：2022-07-06
收稿日期：2022-05-10

摘要：通过研究现有异常网络流量检测技术存在的问题，提出了一种分层自编码器（HAE）集成模型，以无监督的学习方式摆脱了传统检测方法对于样本标签和攻击样本的依赖，以分层集成的方式学习正常流量的多种分布特征提高单个自编码的检测效果。与现有集成学习方式不同，HAE以串行的方式学习上一自编码器学得不好的样本，降低了训练和测试时间。仿真实验结果表明，相比传统的异常检测方法，HAE具有更高的检测率。

关键词：分层自编码器；异常网络流量检测；无监督学习方法；集成学习

Abstract: By studying the problems existing in the existing malicious network traffic detection technology, a hierarchical autoencoder (HAE) ensemble model is proposed, which gets rid of the dependence of traditional detection methods on sample labels and attack samples by unsupervised learning, learns various distribution characteristics of normal traffic by hierarchical integration, and improves the detection effect of single autoencoder. Different from the existing ensemble learning methods, HAE learns the samples that the previous self-encoder does not learn well in a serial way, which reduces the training and testing time. Simulation results show that HAE has a higher detection rate than traditional anomaly detection methods.

Keywords: HAE; malicious network traffic detection; unsupervised learning method; ensemble learning

引用格式：张晓青，谷勇浩，田甜．基于分层自编码器的异常网络流量检测 [J]. 中兴通讯技术，2024, 30(5): 81-86. DOI: 10.12142/ZTETJ.202405012
Citation: ZHANG X Q, GU Y H, TIAN T. Malicious network traffic detection based on hierarchical autoencoder [J]. ZTE technology journal, 2024, 30(5): 81-86. DOI: 10.12142/ZTETJ.202405012

现有的网络攻击种类繁多，随着技术的发展，还会产生新的攻击类型或已有攻击类型的变种，而现有的基于有监督或半监督的恶意软件检测方法只对训练集中出现的恶意软件类别具有较好的检测效果，对于未知类别的恶意软件检测能力较弱。此外，该类方法还存在恶意软件流量获取困难、样本标注代价高等问题。因此，传统的基于有监督或半监督的异常检测方法对训练样本的要求较高，且对未知类别的恶意软件检测能力不足。为此，本文提出了一种新的无监督异常检测方法——分层自编码器（HAE）集成，以分层集成的方式提高自编码器的检测性能。本文的主要贡献如下：

1）基于自编码器提出了一种分层集成的模型HAE，通过学习正常流量的不同分布特征提高单个自编码器的学习效果；

2）基于3个公开数据集，将HAE与其他传统异常检测方法和集成方式进行对比实验，验证了HAE对于网络流量异常检测的有效性。

基金项目：北京邮电大学中央高校基本科研业务费行动计划项目（2021XD-A11-1）；中兴通讯产学研合作项目（HC-CN-20200807013）；广东省信息安全技术重点实验室开放基金项目（2020B1212060078）

1 研究现状

无监督异常检测方法通过学习正常样本的通用特征来区分异常样本，当检测行为与正常行为偏离较大时发出告警信息，不依赖样本标签和异常样本，省去了数据标记、恶意样本捕获等预处理工作，并且对于未知种类的恶意样本具有较好的检测效果，被广泛应用于各个领域，如医疗[1]、军事[2]、交通[3]等。本文从机器学习方法、深度学习方法以及集成学习方法3个方面介绍网络异常流量检测领域的无监督异常检测技术。

常见的机器学习异常检测方法有基于距离或密度的相似度度量方法、基于概率密度的方法等。研究人员[4-5]采用基于K-means聚类方法，将样本间距离作为相似度度量指标寻找正常样本的质心，基于样本到质心的距离计算异常分数，距离越远异常分数越大。高斯混合模型（GMM）是一种基于概率密度的异常检测方法，正常样本比异常样本具有更高的预测概率。BLANCO等[6]对于每维特征使用一个GMM进行建模，通过聚合所有特征的预测概率得到样本的正常分数。无监督机器学习方法实现简单，但是对离群点和噪声点敏感且运算开销大，难以应用于大规模数据集。作为一种有效的无监督降维方法，主成分分析法（PCA）常与其他异常检测方法结合使用进行异常检测[7-8]，以减少计算开销。

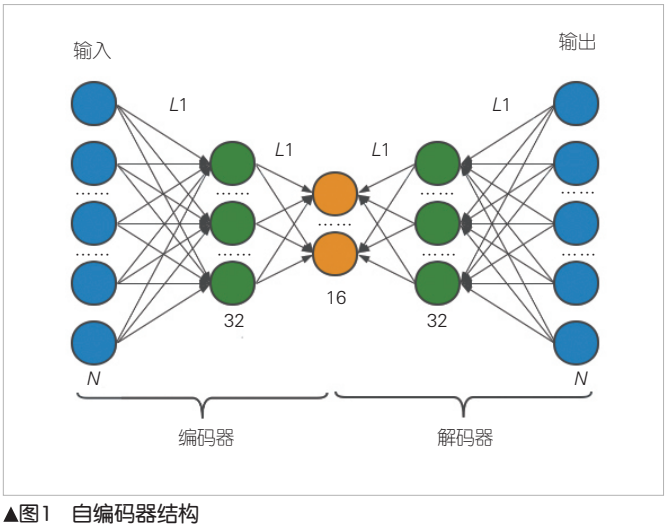
深度学习异常检测方法主要是基于自编码器（AE）的方法和基于生成式对抗网络（GAN）的方法。由于网络流量具有时序性，因此序列自编码器常被用于网络流量异常检测。孙旭日等[9]提出了一种基于长短期记忆（LSTM）的自编码网络结构，由LSTM-Encoder和LSTM-Decoder两部分组成。其中，前者对输入的特征向量时间序列数据进行等长学习表达，后者使用当前的隐含状态和前一步预测的值对编码后的时间序列进行重构，最后结合预设阈值检测异常流量。VU等[10]提出了多分布变分自动编码器（MVAE），在变分自动编码器的损失函数的KL项中引入了数据样本的标记信息，使其更具可分辨性。该标记信息允许MVAE将网络数据样本强制划分到潜在特征空间中不同区域的不同类中，使得网络流量更易区分。杨宏宇等[11]基于变分自动编码器和生成式对抗网络提出了V-G模型：首先使用多簇选择算法（MCFS）对多源数据进行无监督特征选择，然后构建V-G模型进行训练和参数优化，并通过均方根误差函数统计训练过程中的重构误差阈值。

集成学习通过组合多个弱学习器提高单个模型的检测效果，MIRSKY等[12]基于自编码器提出了一种轻量级的异常检测方法Kitsune：首先基于层次聚类将特征空间划分为多个子空间，分别使用一个自编码器进行特征降维与重构，然后使用一个自编码器对重构误差进行非线性聚合，得到样本的异常分数。为了克服自编码器过拟合和泛化能力弱的问题，吴德鹏等[13]使用降噪自编码器进行集成，以一定的概率将输入层节点的值置为零，通过对缺失特征数据进行预测提高自编码器的泛化性能。于振洋等[14]将AdaBoost算法引入网络异常流量检测中，分别对两种弱学习方法进行集成：估计多变量高斯分布和估计超球体区域，首先根据当前训练样本的权重分布得到弱学习器，计算样本的误差信息和弱学习器的权重，然后更新样本的权重分布重新训练得到多个弱学习器，对于测试样本，根据弱学习器的权重组合所有弱学习的输出结果。

2 基于分层自编码器的异常检测方法

2.1 自编码器

自编码器是一种无监督神经网络，主要由编码器和解码器两部分组成。其中，编码器将输入数据x压缩成浅层空间表征z，解码器对浅层空间表征进行重构，通过最小化输入样本x和重构样本$\hat{x}$的均方误差学习输入数据的特征。此外，自编码器作为一种数据压缩算法具有3个显著特征：数据相关、有损压缩、自动学习。常见的自编码器有稀疏自编码器、深度自编码器、卷积自编码、变分自编码器等。综合考虑数据的特点、神经网络的复杂度，文章采用的自编器类型为稀疏自编码器，如图1所示，输入和输出都是维度为N的特征向量，编码器和解码器都是一个两层的多层感知机（MLP），通过添加一个$L1$正则化项对隐层单元施加稀疏性约束，降低模型过拟合的风险。



▲图1 自编码器结构

2.2 集成学习

根据不同的场景，正常流量数据也分为多种，如上传、下载、浏览等。理想情况下，系统希望通过自编码器学习各种类型正常流量的分布特征，但是训练结果往往是模型只在部分训练数据上表现良好，因此可以通过集成学习组合多个自编码器，学习不同场景下正常流量的分布特征，降低误报。常见的集成方法有Bagging、Boosting等，其中Bagging随机抽取训练子集的方式针对性不强，缺乏指导原则，Boosting通过增加预测得不好的样本的权重来学习单个模型学习得不好的样本，这种方式使得它对样本噪声十分敏感。因此针对以上问题，本文提出了一种分层集成的方法，对当前层自编码器没有学好的样本构建下一层自编码器后重新训练，以较小的代价使得模型能够充分学习正常网络流量。

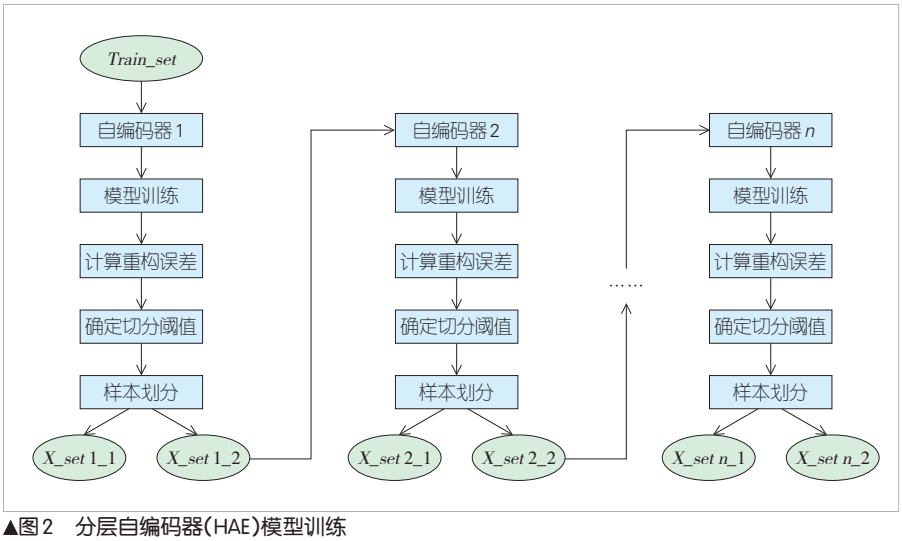
2.3 分层自编码器

为了指导模型学习正常流量的多种分布特征，本文提出分层集成的方式训练多个自编码器，如图2所示。多个自编码器以串行的方式进行集成，不同的自编码器独立训练，训练完成后的自编码器只在部分数据上表现良好。为了使模型学习到另外一部分数据的隐层特征，需要重新构建自编码器对这部分数据进行学习，通过不断地对自编码器表现不好的样本子集构建自编码器重新训练，得到在不同样本子空间上表现良好的自编码器集合。

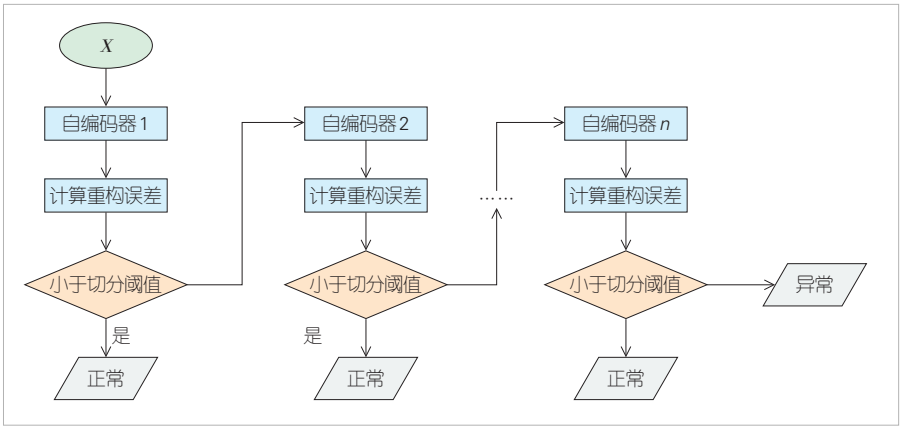
为了划分自编码器表现不好的样本子集，本文将重构误差作为样本集划分的依据，按照一定的比例将重构误差高的样本子集进行重新训练，如图2所示。自编码器训练完成后会根据重构误差将训练数据划分为两部分：X_set1和X_set2。两部分数据子集的大小满足$|X_set1|:|X_set2|=a$，重构误差MSE满足$MSE\{X_set1\}\leqslant b\leqslant MSE\{X_set2\}$，其中$a$是划分比例，$b$是当前自编码器对应的数据切分阈值，为了方便计算，将b设置为$\mathrm{Max}\{MSE\{X_set1\}\}$。$X_set1$的重构误差较低，说明该自编码器学到的隐层特征能够较好表示这一类样本，而X_set2是该自编码器学的不好的样本，将其输入到下一自编码器中重新学习。集成的自编码器的数目n作为模型的超参数，根据验证集的效果进行选取。

HAE模型预测或称如图3所示，每个自编码器相当于一个过滤器，对未知流量进行预测并计算重构误差，若重构误差小于数据切分阈值，则认为符合当前自编码器学习到的样本特征分布，否则输入到下一自编码器进行判断。若未知流量大于所有自编码器的异常阈值，则认为该流量与训练集中正常流量的特征分布不同，将其判定为异常。

本文分析了分层集成和AdaBoost集成的训练和测试复杂度。两种集成都以串行的方式进行集成，并根据上一自编码器的训练结果调整训练数据重新训练。不同的是，AdaBoost集成每次基于整个样本集进行训练，而分层集成则按照一定比例将学得不好的样本进行训练，训练集的规模在逐层递减，训练复杂度更低。在对未知样本进行预测时，AdaBoost集成中所有自编码器都需要参与决策，而分层集成可能在任何一个自编码器判断结束后停止，测试复杂度更低。复杂度



▲图2　分层自编码器(HAE)模型训练



▲图3　分层自编码器(HAE)模型预测

计算结果如表1所示。

3 实验验证与分析

3.1 数据集

为了验证HAE在网络异常流量检测方面的有效性和可用性，本文选择了CIC IDS 2017、UNSWNB15和USTCTFC 3个开源数据集进行对比实验。为了避免数据不均衡对评估指标的计算产生影响，我们选择全部的攻击流量及等量的正常流量作为测试集，其余正常流量作为训练集。不同数据集训练集和测试集规模如表2所示。

CIC IDS 2017数据集[15]基于B-Profile系统[16]模拟正常的主机交互行为，在不同的时间点设置了Brute Force、Heart-bleed、Botnet、拒绝服务（DoS）、分布式拒绝服务（DdoS）、Web Attack、Infiltration共7种攻击场景，包含12种具体的攻击类型，5天捕获了2 425 955条流量记录。我们使用流量特征提取工具CICFlowMeter[17]提取了80维流量特征，包含http、https、安全外壳协议（SSH）、文件传输协议（FTP）、Email共5种应用层协议。

UNSWNB15数据集[18]基于IXIA工具[1]模拟网络环境进行流量采集，分别以不同的攻击频率进行两次场景模拟，耗时31 h捕获了2 540 047条流量记录，包含DoS、Backdoors、Worms、Fuzzers、Analysis、Exploits、Generic、Reconnaissance、Shellcode共9种攻击类型。使用Argus、Bro-IDS两种特征提取工具和12种算法提取了49维流量特征，包含传输控制协议（TCP）、用户数据报协议（UDP）、互联网控制消息协议（ICMP）等多种传输层协议以及http、ftp、ssh、dns等多种服务类型。

USTCTFC数据集[2]中的攻击流量是由CTU大学2011年至2015年从真实网络环境中捕获的，包含Cridex、Geodo、Hitbot、Miuref、Neris、Nsis、Shifu、Tinba、Virut和Zeus共10种攻击类型。正常流量是使用IXIA工具[1]捕获的，包含BitTorrent、Facetime、FTP、Gmail、MySQL、Outlook、Skype、服务器消息块（SMB）、World Of Warcraft和Weibo共10种场景。数据集大小为558 641。

3.2 预处理

CIC IDS 2017数据集和UNSWNB15数据集提供的是特征提取后的流量数据，对于USTCTFC数据集提供的pcap格式的原始报文数据，本文使用特征提取工具CICFlowMeter[17]提取80维流量特征。此外，为了方便后续实验处理，需要对特征提取后的数据进行数据清洗和数据归一化。

1）数据清洗。网络会话捕获不完整或其他网络原因可能会导致特征提取后的个别流量中含有NaN、Inf等异常值，对后续数据处理产生影响。因为这部分流量占比较少，因此本文将包含异常值的流量作为“脏数据”进行剔除。

2）数据归一化。为了消除不同特征维度量纲的影响，本文采用线性函数归一化的方法将各个维度数据的数值范围归一化到0~1之间。归一化的过程如公式（1）所示：

$$X_{\text{norm}} = \frac{X - X_{\min}}{X_{\max} - X_{\min}}, \tag{1}$$

其中，X_{norm}是归一化后的结果，X是原始数据，$X_{\min}$、$X_{\max}$分别是原始数据中的最小值和最大值。

3.3 性能评估指标

为了合理评估模型的性能，本文采用精确率Pre、召回率Rec和F1值作为性能评估指标。精确率又称为查准率，反映了预测结果为异常的样本中真实类别为异常的样本比例；召回率也称查全率，反映了异常样本中预测正确的样本比例；F1值反映了精确率和召回率的综合表现。精确率和

▼表2 数据集攻击类型及规模

数据集	攻击类型	训练集	测试集	数据集大小
CIC IDS 2017	12种	1 764 511	864 444	2 425 955
UNSWNB15	9种	1 897 481	642 566	2 540 047
USTCTFC	10种	127 814	430 827	558 641

▼表1 分层集成和AdaBoost集成的训练和测试复杂度

序号	ABAE	HAE	备注
1	$P = N$	$P = N \times (\frac{1}{1+a})^m$	P为第m个自编码器训练的样本量，N为原始训练集规模，a是HAE样本且分比例。
2	$T = M$	$T = \begin{cases} k, & x\text{为正常流量} \\ M, & \mathbf{x}\text{为异常流量} \end{cases}, k \leqslant M$	T是对未知样本x预测过程中参与决策的自编码器的数目，M为集成的自编码器的总数。

ABAE：基于AdaBoost的集成模型 HAE：分层自编码器

1 http://www.ixiacom.com/products/perfectstorm
2 https://github.com/yungshenglu/USTC-TFC2016

召回率相差较大会导致F1值降低。F1值常被用来衡量模型的整体性能。以上3种指标的计算公式如下：

$$\mathrm{Pre}=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}},\tag{2}$$

$$\mathrm{Rec}=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}},\tag{3}$$

$$\mathrm{F1}=\frac{2\times\mathrm{Pre}\times\mathrm{Rec}}{\mathrm{Pre}+\mathrm{Rec}},\tag{4}$$

其中，TP为恶意流量检测为恶意流量的样本数目，TN为正常流量检测为正常流量的样本数目，FP为正常流量检测为恶意流量的样本数目，FN为恶意流量检测为正常流量的样本数目。

3.4 模型性能评估

为了合理对比不同模型实验效果，本文基于相同的实验环境构建各个模型，选择各个模型最优的测试结果进行对比。为了体现HAE在网络异常流量检测中的优势，本文选择主成分分析方法（PCA）、孤立森林（IForest）和基于频数直方图的异常检测方法（HBOS）进行对比，实验结果如表3所示。其中，各个数据集上HAE的召回率和F1值优势比较明显，说明HAE能够极大提高异常样本的检出率，模型整体性能最优。从精确率指标可以看出，HAE在UNSWNB15和USTCTFC两个数据集上依然具有一定的优势，在CICIDS2017数据集上精确率较低，但是召回率和F1值都很高，说明HAE以较小的代价牺牲精确率，极大地提高了召回率，带来了模型整体性能的提升。综合比较模型的实验结果，HAE相比传统的机器学习方法效果更好，原因主要有两点：一是训练集规模较大，机器学习算法训练困难；二是HAE通过构建多层神经元提取不同层次的数据特征，表征能力更强。

为了验证分层集成的优势，本文选择基于AdaBoost的集成模型（ABAE）、自编码器AE、基于LSTM的时序自编码器（LSTM-AE）[9]进行对比实验，实验结果如表4所示。可以看出，其中各个数据集上HAE和ABAE的F1值均高于单个AE，这说明集成能够提高单个自编码器的学习效果。加入时序特征学习的LSTM-AE模型在CICIDS2017数据集上具有较好的效果，仅次于HAE模型，说明该数据集上时序特征明显，但是在其他两个数据集上效果不佳，甚至不如AE。在CICIDS2017数据集HAE的各项指标均高于ABAE，具有明显优势；在UNSWNB15数据集上HAE的精确率更高说明HAE在该数据集上能够降低误报，牺牲了一定的召回率，导致F1值比ABAE低约0.6%，但是HAE的复杂度比ABAE更低，计算效率更高；在USETCFTC数据集上HAE的召回率和F1值更高，精确率较低说明HAE以较小的代价牺牲了精确率带来了召回率的较大提升和模型整体性能的提高。综合F1值和模型复杂度进行分析，ABAE的效果更好，因为HAE中每一自编码器只专注于学习上一自编码器学得不好的样本，缓解了AdaBoost集成的不足，提升了模型的性能。

4 结束语

为解决现有网络异常检测领域中样本标记难、攻击流量

▼表3　不同数据集上HAE与传统机器学习方法的对比实验结果

数据集	CICIDS2017			UNSWNB15			USTCTFC		
指标	Pre	Rec	F1	Pre	Rec	F1	Pre	Rec	F1
PCA	81.21	43.22	56.42	81.56	44.40	57.50	94.42	71.88	81.62
IForest	80.26	40.64	53.95	78.54	36.32	49.67	94.52	73.35	82.60
HBOS	80.33	40.79	54.11	80.77	41.05	54.44	94.00	66.81	78.11
HAE	68.42	96.25	79.99	97.25	89.47	93.20	96.44	99.27	97.84

HAE：分层自编码器　　Iforest：孤立森林
HBOS：基于频数直方图的异常检测方法　　PCA：主成分分析方法

▼表4　不同数据集上HAE与AE、ABAE的对比实验结果

数据集	CICIDS2017			UNSWNB15			USTCTFC		
指标	Pre	Rec	F1	Pre	Rec	F1	Pre	Rec	F1
AE	69.01	83.67	75.64	90.30	93.28	91.77	95.77	97.07	96.41
ABAE	67.27	91.06	77.38	96.76	90.97	93.78	97.75	95.21	96.46
LSTM-AE	66.20	99.66	79.55	90.18	84.81	91.53	97.62	81.62	88.90
HAE	68.42	96.25	79.99	97.25	89.47	93.20	96.44	99.27	97.84

ABAE：AdaBoost的集成模型　AE：自编码器　HAE：分层自编码器

难以捕获，以及现有集成方式复杂度高、精度低的问题，本文提出了一种轻量有效的无监督集成方式：分层集成，并基于自编码器构建集成异常检测模型HAE，对自编码器学得不好的样本重新训练使得模型能够学习正常数据多种分布特征，提高了模型的检测效果，并通过划分阈值的方式降低模型复杂度。实验结果表明，HAE在多个数据集上优于其他异常检测模型，具有较强的泛化能力和特征学习能力。后续我们将考虑改进样本划分阈值的计算方法，使得模型能够根据训练结果动态调整阈值，并考虑使用变分自编码器学习隐变量的数据分布，降低模型过拟合的风险，提升模型的整体性能。

参考文献

[1] MOHAMED M B, MEDDEB-MAKHLOUF A, FAKHFAKH A. Intrusion cancellation for anomaly detection in healthcare applications [C]//Proceedings of 15th International Wireless Communications & Mobile Computing Conference (IWCMC). IEEE, 2019: 313-318. DOI: 10.1109/IWCMC.2019.8766592

[2] GURUNG S, CHAUHAN S. Performance analysis of black-hole attack mitigation protocols under gray-hole attacks in MANET [J]. Wireless networks, 2019, 25(3): 975-988. DOI: 10.1007/s11276-017-1639-2

[3] GAO L L, LI F, XU X, et al. Intrusion detection system using SOEKS and deep learning for in-vehicle security [J]. Cluster computing, 2019, 22(6): 14721-14729. DOI: 10.1007/s10586-018-2385-7

[4] 贾凡, 严妍, 张家琪. 基于K-means聚类特征消减的网络异常检测 [J]. 清华大学学报(自然科学版), 2018, 58(2): 137-142. DOI: 10.16511/j.cnki.qhdxxb.2018.26.005

[5] LIU H, HAO G, XING B. Entropy clustering-based granular classifiers for network intrusion detection [J]. EURASIP journal on wireless communications and networking, 2020, 2020(1): 4. DOI: 10.1186/s13638-019-1567-1

[6] BLANCO R, MALAGÓN P, BRIONGOS S, et al. Anomaly detection using Gaussian mixture probability model to implement intrusion detection system [M]//Lecture Notes in Computer Science. Cham: Springer International Publishing, 2019: 648-659. DOI: 10.1007/978-3-030-29859-3_55

[7] SALO F, BOU NASSIF A, ESSEX A. Dimensionality reduction with IG-PCA and ensemble classifier for network intrusion detection [J]. Computer networks, 2019, 148: 164-175. DOI: 10.1016/j.comnet.2018.11.010

[8] LIU L Q, XU B, ZHANG X P, et al. An intrusion detection method for Internet of Things based on suppressed fuzzy clustering [J]. EURASIP journal on wireless communications and networking, 2018, 2018(1): 113. DOI: 10.1186/s13638-018-1128-z

[9] 孙旭日, 刘明峰, 程辉, 等. 结合二次特征提取和LSTM-Autoencoder的网络流量异常检测方法 [J]. 北京交通大学学报, 2020, 44(2): 17

[10] VU L, CAO V L, NGUYEN Q U, et al. Learning latent distribution for distinguishing network traffic in intrusion detection system [C]//Proceedings of ICC 2019 - 2019 IEEE International Conference on Communications (ICC). IEEE, 2019: 1-6. DOI: 10.1109/ICC.2019.8762015

[11] 杨宏宇, 王峰岩. 基于无监督多源数据特征解析的网络威胁态势评估 [J]. 通信学报, 2020, 41(2): 143-154

[12] MIRSKY Y, DOITSHMAN T, ELOVICI Y, et al. Kitsune: an ensemble of autoencoders for online network intrusion detection [EB/OL]. [2022-06-15]. http://arxiv.org/abs/1802.09089

[13] 吴德鹏, 柳毅. 基于集成降噪自编码的在线网络入侵检测模型 [J]. 计算机应用研究, 2020, 37(11): 3396-3400. DOI: 10.19734/j.issn.1001-3695.2019.07.0300

[14] 于振洋. 基于Boosting的网络异常流量检测算法研究 [J]. 淮阴工学院学报, 2011, 20(5): 39-43. DOI: 10.3969/j.issn.1009-7961.2011.05.008

[15] SHARAFALDIN I, HABIBI LASHKARI A, GHORBANI A A. Toward generating a new intrusion detection dataset and intrusion traffic characterization [C]//Proceedings of the 4th International Conference on Information Systems Security and Privacy. SCITEPRESS - Science and Technology Publications, 2018: 108-116. DOI: 10.5220/0006639801080116

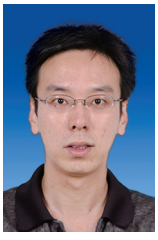
[16] SHARAFALDIN I, GHARIB A, LASHKARI A H, et al. Towards a reliable intrusion detection benchmark dataset [J]. Software networking, 2017, 2017(1): 177-200. DOI: 10.13052/jsn2445-9739.2017.009

[17] LASHKARI A H, DRAPER G G, MAMUN M S I, et al. Characterization of tor traffic using time based features [EB/OL]. (2017-02-20) [2022-06-15]. https://pdfs.semanticscholar.org/d76f/32eb3af1a163c0fde624e9fc229671ca75b6.pdf

[18] MOUSTAFA N, SLAY J. UNSW-NB15: a comprehensive data set for network intrusion detection systems (UNSW-NB15 network data set) [C]//Proceedings of Military Communications and Information Systems Conference (MilCIS). IEEE, 2015: 1-6. DOI: 10.1109/MilCIS.2015.7348942

作 者 简 介

张晓青，北京邮电大学计算机学院在读硕士研究生；主要研究领域为网络安全、数据科学与信息处理。



谷勇浩，北京邮电大学计算机学院硕士研究生导师；主要研究领域为网络安全、数据科学与信息处理；主持参与国家级、省部级及企业资助的科研项目20余项；发表论文近100篇，申请专利20余项。

田甜，中兴通讯股份有限公司资深安全标准与产品总监；研究领域包括APT、M2M安全、WLAN安全、5G安全态势感知等；拥有10余年网络与通信安全领域经验，多次牵头与参与3GPP、IETF、ITU-T、CCSA的安全标准制定工作，拥有数十件欧美授权专利。