基于对比学习的细粒度未知恶意流量分类方法

王一丰[1]，郭渊博[1]，陈庆礼[1]，方晨[1]，林韧昊[2]
（1. 信息工程大学密码工程学院，河南 郑州 450001；2. 郑州大学计算机与人工智能学院，河南 郑州 450001）

摘 要：为了应对层出不穷的未知网络威胁和日益先进的逃逸攻击，针对恶意流量分类问题，提出了一种基于对比学习的细粒度未知恶意网络流量分类方法。所提方法基于变分自编码器，分为已知和未知流量分类 2 个阶段，分别基于交叉熵和重构误差对已知和未知恶意流量分类。与常规方法不同，该方法在各训练阶段中加入了对比学习方法，提高对小样本和未知类恶意流量的分类性能。同时，融合了再训练和重采样等方法，进一步提高对小样本类的分类精度和泛化性能。实验结果表明，所提方法分别提高了对小样本类 20.3%和对未知类恶意类 9.1%的细粒度分类宏平均召回率，并且极大地缓解了部分类上的逃逸攻击。

关键词：网络流量分类；对比学习；变分自编码器；入侵检测
中图分类号：TP393
文献标志码：A
DOI: 10.11959/j.issn.1000−436x.2022180

Method based on contrastive learning for fine-grained unknown malicious traffic classification

WANG Yifeng[1], GUO Yuanbo[1], CHEN Qingli[1], FANG Chen[1], LIN Renhao[2]
1. Department of Cryptogram Engineering, Information Engineering University, Zhengzhou 450001, China
2.School of Computer and Artifical Intelligence, Zhengzhou University, Zhengzhou 450001, China

Abstract: In order to protect against unknown threats and evasion attacks, a new method based on contrastive learning for fine-grained unknown malicious traffic classification was proposed. Specifically, based on variational auto-encoder (CVAE), it included two classification stages, and cross entropy and reconstruction errors were used for known and unknown traffic classification respectively. Different form other methods, contrastive learning was adopted in different classification stages, which significantly improved the classification performance of the few-shot and unknown (zero-shot) classes. Moreover, some techniques (e.g., re-training and re-sample) combined with contrastive learning further improved the classification performance of the few-shot classes and the generalization ability of model. Experimental results indicate that the proposed method has increased the macro recall of few-shot classes by 20.3% and the recall of unknown attacks by 9.1% respectively, and it also has protected against evasion attacks on partial classes to some extent.

Keywords: network traffic classification, contrastive learning, variational auto-encoder, intrusion detection

0 引言

基于流量的网络入侵检测系统（NIDS, network intrusion detection system）是网络空间中最重要的安全设备类型之一，已被广泛应用于各类信息系统的防护中。随着移动互联网和物联网（IoT, Internet of things）等技术的发展，系统的规模不断扩大，且终端设备的种类和数量也不断增多。而这些资源受限的终端设备由于难以部署安全软件或代理且更易存在较多漏洞，更易遭受攻击。这种趋势使 NIDS 在网络防御中的重要性日益凸显。

NIDS 的主要作用是对恶意流量的分类。目前，

收稿日期：2022−06−28；**修回日期**：2022−08−29
基金项目：国家自然科学基金资助项目（No.61501515, No.61601515）
Foundation Item: The National Natural Science Foundation of China (No.61501515, No.61601515)

流量分类方法主要有基于端口、基于载荷和基于流 3 种[1]。然而，如今网络攻击愈发具有对抗性，常采用各种途径规避 NIDS 检测。例如，端口混淆和端口跳变等技术使基于端口的方法准确率大幅下降[2]；加密技术的广泛应用使基于载荷的方法几乎完全失效[3]，此外，隐私和计算开销等问题也是这类方法的缺点[1]。2020 年，超过 70%的恶意活动在通信中使用了加密技术[4]。而基于流的方法不存在上述问题，并且由于机器学习（ML, machine learning）和深度学习（DL, deep learning）的快速发展，这类方法的检测性能大幅提高，现已成为流量分类领域的主流方法[5-6]。然而，以往研究发现这类方法在实际恶意流量检测中存在以下 3 个问题。

1) 机器学习，特别是深度学习的分类性能严重依赖于训练数据的数量和质量[7]。但由于实际中标注数据成本昂贵或存在新型网络威胁，往往没有足够标注数据可用。此外，更细粒度的分类能提供更多威胁信息，有助于安全专家快速响应。而以往大多研究没有考虑此问题，缺乏在细粒度小样本（C2FS, coarse-to-fine few-shot）条件下的评估结果。

2) 恶意流量检测也是一个开集分类问题，数量快速增长的未知攻击（如恶意软件新变种、零日攻击以及针对 IoT 等新兴技术的攻击）产生了各类未知恶意流量，现今的 NIDS 需应对未知恶意流量分类的挑战[8]。而以往大多实验结果是在闭集上，未考虑对未知类的检测。

3) 机器学习算法容易遭受逃逸攻击，或者说对抗样本攻击[9]。研究表明，某些恶意流量只需添加不影响恶意功能的微小扰动即可绕过 NIDS 的检测[10-11]。此外，流量欺骗[12]、混淆[13]等技术的发展也使恶意类流量可以伪装接近于良性类，从而误导分类器[14]。

为了解决以上问题，本文提出一种基于对比学习[15]的细粒度未知恶意流量分类方法，主要贡献如下。

1) 提出了在 C2FS 条件下的恶意流量检测任务，并提出了一种基于条件变分编码器（CVAE, conditional variational auto-encoder）[16]和极值理论（EVT, extreme value theory）[17]的半监督在线学习方法。相比以往方法，本文在一个框架下实现了对已知、未知和小样本类上的细粒度高性能分类。

2) 在流量分类的各阶段设计并加入了对比损失，并采用了再训练架构，使对小样本恶意类流量的分类性能和模型的泛化性能相比以往大幅提高。

3) 在公开数据集上的实验结果表明，相比以往方法，本文方法对所有恶意类，特别是小样本类的分类精度更高，且能够有效降低逃逸攻击效果。

1　相关工作

1.1　基于机器学习的恶意流量分类

恶意流量分类的基本目标是区分良性类和恶意类流量。随着数据规模的增加，相比传统方法，机器学习在恶意流量分类中的性能优势愈发凸显[18]。采用机器学习的方法具有使用场景广、分类准确率高、能分类加密流量等优点。这些研究的区别主要在于流特征、分类对象和分类算法的选择。常用的流特征包括头部、负载、时空和统计特征[7]。依据颗粒度，可将分类对象划分为分组、主机和会话级[19]。常用的分类算法包括支持向量机（SVM, support vector machine）[20]、高斯混合模型（GMM, Gaussian mixture model）[21]、随机森林（RF, random forest）[22]、卷积神经网络（CNN, convolutional neural network）[23]等，且都在实验环境下取得了不错的效果。

基于头部和负载特征的方法在应对精心设计的攻击时表现不佳。例如，各类公共服务（社交网站、云平台等）由于允许用户定义内容，常被用作隐蔽信道发布命令或传递窃取信息[24]。加密或随机填充等方式也会改变负载特征，从而导致分类性能下降。相比之下，基于统计和时空特征的方法虽然也存在上述问题，但由于在保持恶意功能条件下修改特征成本较大且往往会产生新的可分类特征[13]，因此更适用于具有对抗性的恶意流量分类。

部分研究还考虑了实际应用中细粒度、小/零样本等现实需求和问题。文献[25]提出了一种多级半监督学习框架来缓解恶意流量分类中的类不平衡问题。文献[8]采用 CVAE 和 EVT 来实现流量的细粒度检测和对未知类的分类。但据本文所知，目前缺乏同时考虑上述 3 个问题的恶意流量分类方法。

1.2　开集分类与流量分类

开集分类是指模型通过在训练阶段学习已知类样本，并在测试阶段辨别一个样本是已知类还是未知类。目标是找到一个可测函数 f（$f(x)>0$ 表示分类正确），既要最小化已知分类经验风险 $R_{\varepsilon}\left(f(V)\right)$，也要最小化开放空间风险 $R_o(f)$ [26]，即

$$\arg\min_{f\in H}\left\{R_o(f)+\lambda_r R_{\varepsilon}\left(f(V)\right)\right\} \tag{1}$$

其中，V 是训练数据，λ_r 是正则化常数。

开集分类包括判别和生成两类方法。在判别方法中，实现开集分类主要有基于稀疏表示[27]、极值机（EVM, extreme value machine）[28]、OpenMax[29]等。在生成方法中，主要有基于实例生成[30]和非实例生成[31]两类。在这些方法中，EVT 最常被引入用以对未知类进行评估。当前流量分类中考虑开集分类问题的文献还相对较少。文献[32]采用 Weibull-calibrated SVM 对流量进行细粒度分类，并采用 EVT 评估校准。文献[33]通过 EVT 近似计算每个已知类的边缘距离分布，实现开集流量分类。

1.3 恶意流量分类中的逃逸攻击

逃逸攻击（或对抗样本攻击）是指攻击者在不改变目标系统的情况下，通过构造特定对抗样本以欺骗目标系统的攻击。机器学习，特别是深度学习方法普遍存在容易遭受逃逸攻击的问题[34]，且各类对抗样本生成方法层出不穷，如 FGSM（fast gradient sign method）[35]、FFF（fast feature fool）[36]等。

逃逸攻击近年来已应用于恶意流量伪装中，以逃避基于机器学习的 NIDS 检测。文献[37]采用生成对抗网络（GAN, generative adversarial network）修改流量的非功能性特征，将原始恶意流量转换为对抗流量绕过黑盒 NIDS 检测。文献[38]同样采用 GAN 模型学习良性流特征指导对抗样本生成，以传出/传入数据包等 6 个特征实现逃逸攻击。文献[10]不仅实现了基于 GAN 的加密恶意流量伪装，还模仿了良性类的主机级通信时间特征。随着逃逸攻击在恶意流量分类中不断发展，未来 NIDS 中需要考虑对此类攻击的防御。

1.4 对比学习

对比学习属于表征学习，目的是学习一种数据变换方式，其更容易解决下游任务。通过学习数据的特征表示，使特征空间中相同类数据较近，不同类数据彼此远离。大多数方法是基于对比损失实现的。而其他方法，如 BYOL（bootstrap your own latent）模型[39]虽然没有采用负样本，但其多层感知机（MLP, multilayer perception）预测器也可以视作负样本网络。文献[40]指出对比学习模型的性能与负样本的数量和质量相关。本文总结了当前 3 种主流的对比学习方法。

1) 以 SimCLR[41]为代表的方法。这类方法将当前训练批次中的其他类样本作为负样本，通过对比损失实现对比学习。这类方法训练难度较大且会丢失以往部分负样本。

2) 以 MoCo[42]为代表的方法。这类方法维护一个大的先入先出负样本队列（$x_{-1}, x_{-2}, \cdots, x_{-m}$），每次训练更新最旧的一小批负样本。经典的 MoCo 架构如图 1 所示，采用模板网络 f_M 来实现对正负样本的特征提取。对于 f_M，其初始化参数为原始特征提取网络 f_q，训练时基于动量缓慢更新参数。这类方法因不需要反向传播而计算量较小，但负样本更新速度较慢。

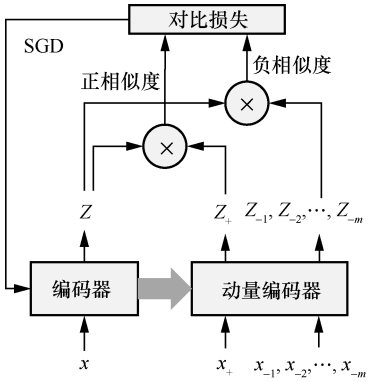


图 1 经典的 MoCo 架构

3) 以 AdCo[43]为代表的方法。这类方法设计负样本网络表示“整体”负样本，用以生成高质量负样本信息[44]。这类方法由于需要设计训练负样本网络，模型结构比较复杂。

2 方法设计

现有基于流特征的 NIDS 存在缺乏标注数据和易受逃逸攻击的问题，本节在文献[8]的基础上，受文献[45-46]启发将对比学习与 CVAE 结合，以实现已知、未知、小样本类的流量细粒度分类。

流量的流特征可以采用 CICFlowMeter[47]等流量特征提取工具提取。本文中流量的流特征经过提取和预处理后记作 d 维实值特征向量 $x_i \in R^d$，对应类标签记作 $y_i \in \{0,1\}^{k+f+1}$，用 $k+f+1$ 维的独热（one-hot）向量表示。记 O 为 one-hot 编码函数，作用是将输入向量的最大维度设为 1，其余维度值设为 0。类标签集合可以表示为 $C_{\text{all}} = \{B, M_1, \cdots, M_k, M_{k+1}, \cdots, M_{k+f}, M_{k+f+1}\}$，其中，$B$ 表示良性类，$M_1, \cdots, M_k$ 表示 k 类大样本已知恶意类，$M_{k+1}, \cdots, M_{k+f}$ 表示 f 类小样本已知恶意类，M_{k+f+1} 表示未知恶意类。训练集 $S_{\text{train}} = S_{\text{train}-B} \cup S_{\text{train}-K} \cup S_{\text{train}-F}$ 由大量良性类样本 $S_{\text{train}-B} = \{(x_i, y_i) \mid y_i = B\}$、一定量的已知恶意类样

本 $S_{\text{train}-K}=\{(x_i,y_i)\mid y_i\in\{M_1,\cdots,M_k\}\}$ 以及极少量的小样本已知恶意类样本 $S_{\text{train}-F}=\{(x_i,y_i)\mid y_i\in\{M_{k+1},\cdots,M_{k+f}\}\}$ 组成。本文利用训练集 S_{train} 训练分类网络 P_θ，使其对所有类样本 $x\in\{x_i\mid y_i\in C_{\text{all}}\}$ 的预测结果都尽量接近其真实标签 y，即目标为

$$\max_{(x,y)\in S_{\text{train}}}\log P_\theta(y\mid x) \tag{2}$$

本文提出了一个基于流特征的流量分类方法，可同时实现对已知、未知、小样本类的流量细粒度分类，整体结构如图 2 所示。模型分为已知恶意流量训练、未知恶意流量训练和测试 3 个阶段。首先，在已知恶意流量训练阶段采用 CVAE 模型，并设计对比损失结合 CVAE 编码器训练，使不同类样本在潜特征空间中进一步区分。相比以往方法，加入对比损失后对于小样本类检测有显著提升。其次，在模型未知恶意流量训练阶段采用 EVT 模型。当模型在已知恶意流量训练后，固定编码器和解码器的网络参数训练重构网络参数，并在重构阶段也加入了对比损失，基于重构误差训练 EVT 模型，使模型对未知类的分类性能显著提高。最后，在测试阶段用训练后的 CVAE 和 EVT 模型对新样本进行分类。并且重采样、再训练和 L2 正则化等

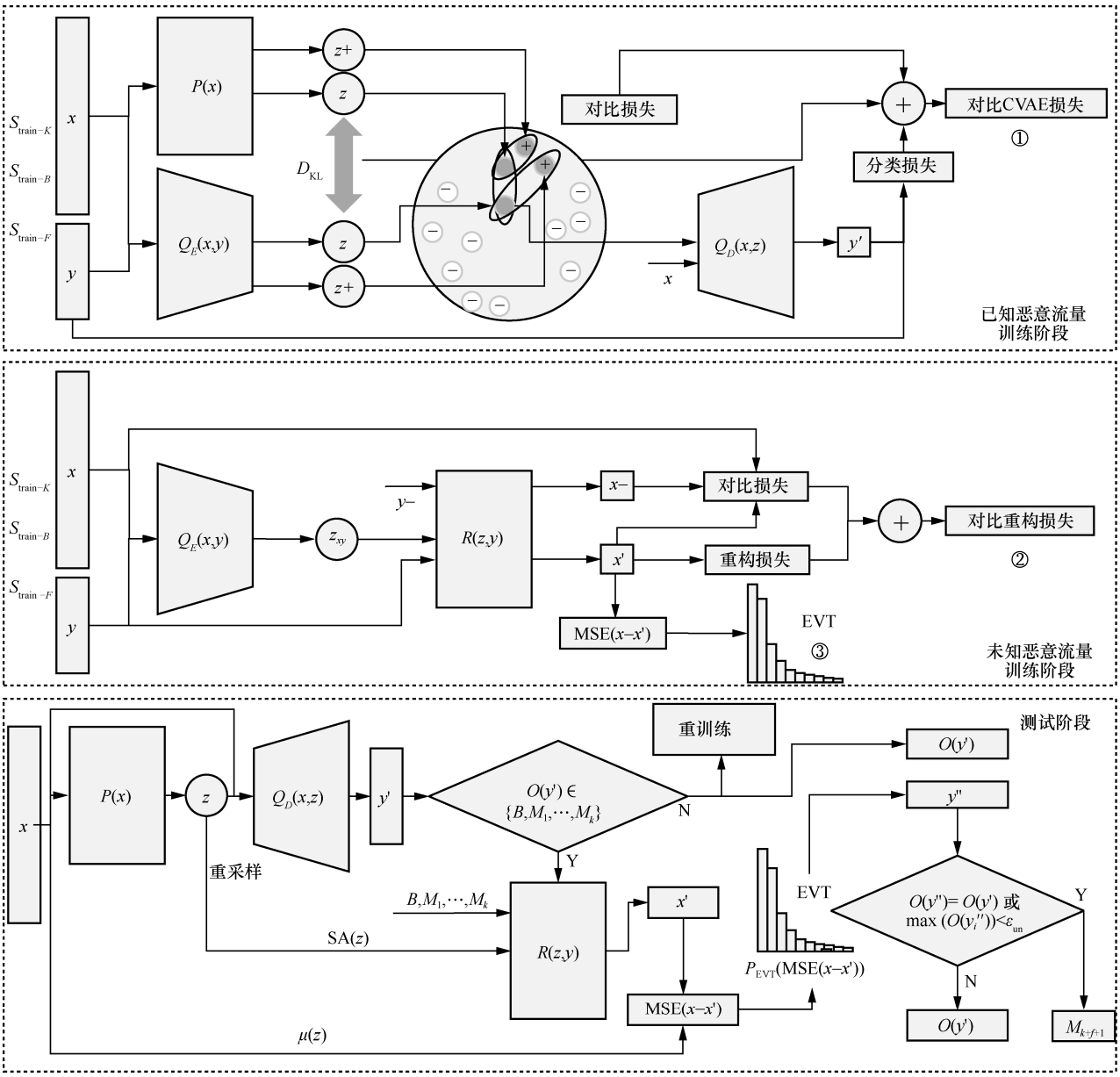


图 2　基于对比学习的细粒度恶意流量检测模型整体结构

技巧在此也被采用，以进一步提升模型的检测和泛化性能。

2.1 已知恶意流量训练阶段

该阶段结构如图 2 中的已知恶意流量训练阶段所示，目标是采用训练集 S_{train} 对良性类（B）和已知恶意类（$M_1,\cdots,M_{k+f}$）样本进行分类。本文在此阶段采用对比学习结合 CVAE 模型[8]实现。CVAE 可视为有监督 VAE 模型，在恶意流量检测中表现优异，更重要的是其独特的“解码重构”架构可以帮助分类未知恶意类样本。

首先，将样本 $x\in R^d$ 转化为更低维的潜特征 $z\in R^h, h<<d$，更低维的特征空间计算速度更快且能更好地区分不同类的样本[8]。对于给定 x,y，其 z 的真实分布记为 $P(z\mid x,y)$，采用一个网络 $Q(z\mid x,y)$ 近似模拟 $P(z\mid x,y)$。对于目标式(2)的 $\log P_\theta(y\mid x)$ 可以改写为[16]

$$\begin{aligned}\log P_\theta(y\mid x)=&\int Q(z\mid x,y)\log P(y\mid x,z)\mathrm{d}z-\\&D_{\text{KL}}\left(Q(z\mid x,y)\,\|\,P(z\mid x)\right)+\\&D_{\text{KL}}\left(Q(z\mid x,y)\,\|\,P(z\mid x,y)\right)\end{aligned}\tag{3}$$

其中，D_{KL} 表示 KL 散度[48]，用于衡量 2 个分布之间的差异。式(3)中前两项合称为 CVAE 证据下界值（ELBO, evidence lower-bound），即

$$\begin{aligned}L_{\text{ELBO}}=&\int Q(z\mid x,y)\log P(y\mid x,z)\mathrm{d}z-\\&D_{\text{KL}}\left(Q(z\mid x,y)\,\|\,P(z\mid x)\right)\end{aligned}\tag{4}$$

由于式(3)第三项中 $P(z\mid x,y)$ 无法计算，但 KL 散度一定非负，因此 $\log P_\theta(y\mid x)\geqslant L_{\text{ELBO}}$，则式(2)可以等价为

$$\max_{(x_i,y_i)\in S_{\text{train}}} L_{\text{ELBO}}\tag{5}$$

其中，$P(z\mid x)$、$Q(z\mid x,y)$、$P(y\mid x,z)$ 均采用 GMLP（Gaussian multi-layered perceptron）近似模拟。假设 $P(z\mid x)$ 服从正态分布，采用先验网络 $P(x)$ 近似；$Q(z\mid x,y)$ 采用编码网络 $Q_E(x,y)$ 近似，对应 CVAE 编码器；$P(y\mid x,z)$ 采用解码网络 $Q_D(x,z)$ 近似，对应 CVAE 解码器。则基于目标式(5)，$P(x)$、$Q_E(x,y)$ 和 $Q_D(x,z)$ 采用随机梯度下降（SGD, stochastic gradient descent）算法通过最小化损失函数式(6)更新网络参数，实现流量分类

$$\begin{aligned}L_{\text{CVAE}}=&-\mathbb{E}_{S_{\text{train}}}\left(y\log Q_D(x,z)+(1-y)\log\left(1-Q_D(x,z)\right)\right)+\\&D_{\text{KL}}\left(Q_E(x,y)\,\|\,P(x)\right)\end{aligned}\tag{6}$$

其中，$z=\text{SA}\left(Q_E(x,y)\right)$ 是 $Q_E(x,y)$ 编码后正态分布的采样，SA 表示采样函数。

但以上 CVAE 模型存在一定缺陷：一是难以分类小样本类，二是易遭受对抗样本的攻击。为了缓解这 2 个问题，本文学习潜特征 z 时加入了对比学习。在潜特征空间中，本文希望同类样本之间更加紧凑，异类样本之间更加分散。这样即使是小样本类也可与其他类区分开，并且有助于分类未知恶意类流量和防御逃逸攻击。

为了实现对比学习，对于任意样本对 $(x_i,y_i)\in S_{\text{train}}$，$P(x_i)$ 首先应与同类样本 x_+ 的 $P(x_+)$ 接近，与异类样本 x_- 的 $P(x_-)$ 远离；同样 $Q_E(x_i,y_i)$ 应与同类样本 x_+ 的 $Q_E(x_+,y_i)$ 接近，与异类样本 x_- 的 $Q_E(x_-,y_i)$ 远离；最后，$P(x_i)$ 应与其对应的 $Q_E(x_i,y_i)$ 接近，与其他不正确的 $y_-\neq y_i$ 生成的 $Q_E(x_i,y_-)$ 远离。而对比学习的损失函数可以采用 InfoNCE 损失[49]的形式来表示

$$\begin{aligned}L_{\text{Cont}}=&-\mathbb{E}\left[\log\frac{\exp\left(\dfrac{P(x_i)P(x_+)}{\tau}\right)}{\sum\exp\left(\dfrac{P(x_i)P(x_-)}{\tau}\right)}\right]-\\&\mathbb{E}\left[\log\frac{\exp\left(\dfrac{Q_E(x_i,y_i)Q_E(x_+,y_i)}{\tau}\right)}{\sum\exp\left(\dfrac{Q_E(x_i,y_i)Q_E(x_-,y_i)}{\tau}\right)}\right]-\\&\mathbb{E}\left[\log\frac{\exp\left(\dfrac{P(x_i)Q_E(x_+,y_i)}{\tau}\right)}{\sum\exp\left(\dfrac{P(x_i)Q_E(x_i,y_-)}{\tau}\right)}\right]\end{aligned}\tag{7}$$

其中，τ 是归一化超参数。计算时对 $P(x)$ 和 $Q_E(x,y)$ 进行重采样，以 $\text{SA}\left(P(x)\right)$ 和 $\text{SA}\left(Q_E(x,y)\right)$ 计算损失函数。

对比学习中负样本的选择至关重要。为了兼顾效率与精度，本文对不同类型 x 采用不同策略选取负样本。首先，采用 SimCLR 类方法迅速训练 $P(x)$、$Q_E(x,y)$；当模型收敛速度下降后，选择负样本时采用 AdCo 类思想，选择针对性负样本。即当样本对 $(x_i,y_i)\in S_{\text{train}-K}$ 时，选择属于同一粗粒度类但不同细粒度类或良性类样本作为负样本；最后，对于小样本类样本对 $(x_i,y_i)\in S_{\text{train}-F}$，希望其与尽可能多的负样本训练，因此对于这部分样本采用 MoCo 类方法训练至模型收敛。

本文在 CVAE 模型训练时加入了对比损失，以不平衡采样数据训练整体的 CVAE 模型，其损失函数如式(8)所示，对应图 2 中的①。

$$L_{\text{All}} = L_{\text{CVAE}} + \lambda_C L_{\text{Cont}} \tag{8}$$

其中，λ_C 是权重超参数。

该阶段的训练策略如算法 1 的步骤 1)～步骤 16)所示。

2.2　未知恶意流量训练阶段

该阶段结构如图 2 中的未知恶意流量训练阶段所示，目标是对未知类（M_{k+f+1}）恶意样本进行分类。在此阶段，本文采用对比学习结合 CVAE 重构网络和 EVT 方法实现流量分类。

在 CVAE 模型完成 2.1 节训练后，首先固定 $Q_E(x,y)$ 的参数训练重构网络 $R(z,y)$。重构网络 $R(z,y)$ 用于对输入潜特征 z（由编码器网络 Q_E 生成）和给定标签 y' 并生成重构样本 x'。其目标是编码网络 Q_E 输入的 x 和解码网络 Q_D 输出的 x' 尽量接近。重构损失函数如式(9)所示，重构损失采用均方误差（MSE, mean square error）来衡量。

$$L_{\text{Recon}} = \mathbb{E}_{S_{\text{train}}} \text{MSE}\left(x_i, R\left(z_i, y_i\right)\right) \tag{9}$$

对于未知类检测思路是训练后的 CVAE 模型对已知类样本对 $\left(x_i, y_i\right) \in \left\{\left(x_i, y_i\right) \mid y_i \in \left\{M_1, \cdots, M_{k+f}\right\}\right\}$ 的预测标签 y_i' 大概率是正确的，即 $y_i' = y_i$，其重构样本 $x_i' = R\left(z_i, y_i'\right)$ 也应接近 x_i；而对未知类样本对 $\left(x_i, y_i\right) \in \left\{\left(x_i, y_i\right) \mid y_i = M_{k+f+1}\right\}$，其预测类标签 y_i' 必定错误，即 $y_i' \neq y_i$，其重构样本 x_i' 应与原样本 x_i 相差较大。这样就可以判断新样本 x_i' 是否属于未知恶意类。

为了减少对未知类的误判，训练时重构样本 x_i' 应与输入样本 x_i 接近，与其他类样本 $y\_ \neq y_i$ 生成的 $R\left(z_i, y\_\right)$ 较远。为此，采用对比学习式(10)放大这种需求。

$$L_{\text{ContRecon}} = -\mathbb{E}\left[\log \frac{\exp\left(\dfrac{R(z_i, y_i)x_i}{\tau}\right)}{\sum_{j=0}^{k+f} \exp\left(\dfrac{R(z_i, y_i)R(z_i, y_j)}{\tau}\right)}\right] \tag{10}$$

最后利用式(11)训练重构网络 $R(z,y)$。训练时固定 $P(x)$ 和 $Q_E\left(x,y\right)$ 参数，实际计算时 $z = \text{SA}\left(Q_E(x,y)\right)$ 由 $Q_E(x,y)$ 编码后正态分布采样，对应图 2 中的②。由于对比损失几何上类似余弦相似度，因此在重构样本时潜特征 z 也需归一化。

$$L_{\text{All}} = L_{\text{Recon}} + \lambda_R L_{\text{ContRecon}} \tag{11}$$

其中，λ_R 是权重超参数。

未知恶意类分类时，由于模型对不同已知类的分类和重构能力不同，难以采用统一阈值进行分类。为此，本文采用 EVT 为每个已知大样本类估计分类阈值。EVT 认为对于任意的随机变量 X，其极值相对于阈值 t 超出的部分应服从 GPD（generalized pareto distribution）分布

$$\text{P}(X - t > x \mid X > t) \sim \left(1 + \frac{\gamma x}{\sigma}\right)^{-\frac{1}{\gamma}} \tag{12}$$

其中，$\gamma, \sigma > 0$，实际中可以采用极大似然估计方法计算 γ, σ，对数似然函数表示为

$$\log L(\gamma, \sigma) = -N_t \log \sigma - \left(1 + \frac{1}{\gamma}\right) \sum_{i=1}^{N_t} \left(1 + \frac{\gamma\left(X_i - t\right)}{\sigma}\right) \tag{13}$$

其中，N_t 表示观测数据中超过阈值 t 的样本数量，X_i 表示观察数据。本文采用 $S_{\text{train}-B} \bigcup S_{\text{train}-K}$ 中样本的重构误差作为观察数据，分别对大样本已知类（$B, M_1, \cdots, M_k$）构建 k+1 个 EVT 模型，对应于图 2 中的③。

该阶段的训练策略如算法 1 的步骤 17)～步骤 24)所示。

2.3　细粒度恶意流量的训练策略与测试阶段

本文模型的整体训练策略如算法 1 所示。

算法 1　细粒度恶意流量分类模型训练算法

输入　标记训练数据集 S_{train}

输出　训练后的 $P(x)$、$Q_E(x,y)$、$Q_D(x,z)$ 和 $R(z,y)$，以及 k+1 个训练后的 EVT 模型参数 $\left\{t_i, \gamma_i, \sigma_i\right\}_{i=0}^{k}$

1) 随机初始化 $P(x)$、$Q_E(x,y)$、$Q_D(x,z)$、$R(z,y)$；
2) for $\left(x_i, y_i\right) \in S_{\text{train}}$ do
3)　　计算 $Q_E\left(x_i, y_i\right)$ 和 $P\left(x_i\right)$ 得到潜特征；
4)　　随机选择负样本，基于式(8)计算损失和梯度，并采用 SGD 更新 $P(x)$、$Q_E(x,y)$、$Q_D(x,z)$ 参数；
5) end for
6) 重复步骤 2)～步骤 5)，直至模型大致收敛；
7) for $\left(x_i, y_i\right) \in S_{\text{train}-K}$ do
8)　　计算 $Q_E\left(x_i, y_i\right)$ 和 $P\left(x_i\right)$ 得到潜特征；
9)　　选择 x_i 的同一粗粒度类但不同细粒度类

或良性类样本为负样本，基于式(8)采用 SGD 更新 $P(x)$、$Q_E(x,y)$、$Q_D(x,z)$ 参数；

10) end for

11) 重复步骤 7)～步骤 10)，直至模型大致收敛；

12) for $(x_i,y_i)\in S_{\text{train}-F}$ do

13) 计算 $Q_E(x_i,y_i)$ 和 $P(x_i)$ 得到潜特征；

14) 采用 MoCo 架构，维持大量负样本，基于式(8)采用动量方式更新 $P(x)$、$Q_E(x,y)$、$Q_D(x,z)$ 参数；

15) end for

16) 重复步骤 12)～步骤 15)，直至模型收敛；

17) for $(x_i,y_i)\in S_{\text{train}-B}\bigcup S_{\text{train}-K}$ do

18) 计算 $Q_E(x_i,y_i)$ 得到潜特征；

19) 固定当前 $P(x)$、$Q_E(x,y)$、$Q_D(x,z)$ 网络参数，基于式(11)采用 SGD 更新 $R(z,y)$ 参数；

20) end for

21) 重复步骤 17)～步骤 20)，直至模型收敛；

22) for $c_i\in\{B,M_1,\cdots,M_k\}$ do

23) 计算 c_i 类所有训练样本的重构误差，并将其作为观察数据，选取经验阈值 t_i 并基于式(13)估算 c_i 类的 EVT 模型参数 γ_i,σ_i；

24) end for

训练完成后，模型在测试阶段的流程如图 2 中的测试阶段所示。具体来说，对于一个待分类的新样本 x_i，首先采用先验网络 P 得到潜特征分布 z_i。接着解码器 Q_D 给出预测标签 $y_i'=Q_D(x_i,\text{SA}(z_i))$。若 $O(y_i')\in\{M_{k+1},\cdots,M_{k+f}\}$ 为小样本恶意类，则直接将 $O(y_i')$ 作为分类向量输出；若 $O(y_i')\in\{B,M_1,\cdots,M_k\}$ 为良性类或大样本已知恶意类，则继续用重构网络 R 计算 x 与所有大样本已知类的标签计算重构样本 $R(\text{SA}(z_i),\{B,M_1,\cdots,M_k\})$ 及其重构误差 $\text{MSE}(x_i-x_i')$。之所以不对被分类为小样本类的样本进行重构，是因为模型对小样本类分类能力不足。接着，基于大样本已知类 EVT 模型，对当前重构误差计算概率 $P_{\text{EVT}}(\text{MSE}(x_i-x_i'))$，用于对 y_i' 的修正，得到修正后的概率向量 y_i''。若分类结果改变（即 $O(y_i'')\neq O(y_i')$）或其最大值小于阈值（即 $\max(O(y_i''))<\varepsilon_{\text{un}}$），则最终分类结果为未知恶意类 M_{k+f+1}；否则，分类结果仍为 $O(y_i'')$。

为了进一步提升对小样本类的分类能力，本文还采用半监督学习中的再训练架构。测试时的无标记样本 x_i 经过模型预测得到分类向量 y_i'，若 $O(y_i')\in\{M_{k+1},\cdots,M_{k+f}\}$，则先基于式(14)分析 y_i' 质量并筛选其中高质量伪标签 $\tilde{y}_i'$。其中高置信度样本对 $(x_i,\tilde{y}_i')$ 作为新标注数据，再次重新用于模型训练以提升对小样本类的分类能力。再训练时只更新解码器 $Q_D(x,z)$ 参数。

$$\tilde{y}_i'=\begin{cases}O(y_i'), & H(y_i')<\varepsilon_{\text{re}}\\ \text{ignore}\end{cases}\tag{14}$$

其中，$H(y)=-\sum_{c=0}^{k+f}p_c\log p_c$，$p_c$ 表示 $\text{soft}\max(y)$ 向量的第 c 个值，ε_{re} 为预先定义的交叉熵阈值。

对抗逃逸攻击需要增强模型的泛化能力[50]，常见方法主要分为两类：一是对数据处理，例如通过添加随机噪声进行数据增强、部署检测器分类输入数据、中间层随机 Dropout 等[51]；二是对模型参数处理，例如 L1、L2 正则化[52]和对抗训练[53]等。

本文采用了 3 种技巧来增强模型的泛化能力以降低逃逸攻击的效果。首先，训练和测试时在潜特征空间中进行重采样，相当于对数据添加了噪声进行数据增强，而这其中对比损失的加入使重采样对模型原本分类性能的影响减少。其次，模型在训练损失函数式(8)和式(11)时加入了 L2 正则项。文献[35]指出 L2 正则化能有效增强线性模型的泛化能力且容易实现。最后，逃逸攻击在未知类分类阶段大多被分类为未知攻击，实现对抗样本的分类。此外，对比学习在此也提高了模型的泛化性能。

3 实验分析

本节在一个恶意流量分类文献中广泛采用的公开数据集 NSL-KDD[54]上评估了所提方法。首先，对所提方法各个组件的有效性进行了分析。其次，将所提方法与现有方法进行了对比分析。实验表明，所提方法在分类小样本恶意类、分类未知恶意类和抵御逃避攻击等方面明显优于其他方法。

3.1 数据集

本文实验采用 NSL-KDD[54]数据集，该数据集是著名 KDD 99 数据集的修订版本。NSL-KDD 数据集包括良性类和 39 种细粒度攻击类。其中，攻击可以分为 4 个粗粒度类，即拒绝服务（DoS, denial

of service）类攻击、扫描（Probe）类攻击、远程入侵（R2L, remote to login）类攻击和本地提权（U2R, user to root）类攻击，具体数据分布如表 1 所示。本文选择 NSL-KDD 作为实验数据集主要基于以下两点：1) NSL-KDD 虽然发布时间较早，但一直被高水平研究所采用，可以与经典和先进方法比较；2) 更重要的是，相比其他公开数据集，NSL-KDD 包含了丰富的细粒度类标签，适用于本文场景，而其他数据集大多只有粗粒度标签，或细粒度标签种类不够丰富。

结合表 1 的数据分布和实际情况发现，DoS 类和 Probe 类在实际中如果存在，则一般有大量标记样本。而 U2R 类和 R2L 类在实际中存在较少，更频繁存在小样本情况。本文在表 1 中列举并采取了接近实际的实验设置。对于小样本类，每类只选取 5 个样本（5-shot）用于训练。

表 1　　NSL-KDD 数据集各细粒度类的数据分布

粗粒度类	细粒度类	训练集数量	测试集数量	本文类别	本文训练采用数量
良性类（benign）	良性类（benign）	67 343	9 711	已知	67 343
DoS 类	apache2	0	737	未知	0
	back	956	359	已知	956
	land	18	7	未知	0
	neptune	41 214	4 657	已知	41 214
	mailbomb	0	293	未知	0
	pod	201	41	未知	0
	processtable	0	685	未知	0
	smurf	2 646	665	已知	2 646
	teardrop	892	12	未知	0
	udpstorm	0	2	未知	0
Probe 类	ipsweep	3 599	141	已知	3 599
	mscan	0	996	未知	0
	nmap	1 493	73	未知	0
	portsweep	2 931	157	已知	2 931
	saint	0	319	未知	0
	satan	3 633	735	已知	3 633
U2R 类	buffer_overflow	30	20	小样本	5
	httptunnel	0	133	未知	0
	loadmodule	9	2	未知	0
	perl	3	2	未知	0
	ps	0	15	未知	0
	xterm	0	13	未知	0
	rootkit	10	13	小样本	5
	sqlattack	0	2	未知	0
R2L 类	worm	0	2	未知	0
	ftp_write	8	3	未知	0
	guess_passwd	53	1 231	已知	53
	imap	11	1	未知	0
	multihop	7	18	小样本	5
	named	0	17	未知	0
	phf	4	2	未知	0
	sendmail	0	14	未知	0
	snmpgetattack	0	178	未知	0
	snmpguess	0	331	未知	0
	spy	2	0	未知	0
	warezclient	890	0	已知	890
	warezmaster	20	944	小样本	5
	xsnoop	0	4	未知	0
	xlock	0	9	未知	0

3.2 评估标准

在分类任务中，TP 为被模型预测为正类的正样本，TN 为被模型预测为负类的负样本，FP 为被模型预测为正类的负样本，FN 为被模型预测为负类的正样本，则分类任务通常采用以下 3 个衡量指标：精度（Precision）、召回率（Recall）和 F_1 值。

$$\text{Percision}=\frac{\text{TP}}{\text{TP}+\text{FP}} \tag{15}$$

$$\text{Recall}=\frac{\text{TP}}{\text{TP}+\text{FN}} \tag{16}$$

$$F_1=\frac{2\text{PercisionRecall}}{\text{Percision+Recall}} \tag{17}$$

在已知类分类阶段，本文同样关注稀疏类（小样本类），因此在训练时采用宏（macro）平均值对模型分类性能进行评估，式(18)表示宏平均 F_1 值。在后续模型性能评估时，都采用宏平均指标。

$$F_{1-\text{macro}}=\frac{1}{k+f+1}\sum_{c_k\in\{B,\cdots,M_{k+f}\}}F_{1-c_k} \tag{18}$$

3.3 有效性分析

本节对本文模型中各个组件的有效性进行分析。实验选择了 200 次训练后的各模型进行比较分析。

本节首先讨论了第 2 节中几个重要超参数的设置。式(8)中 λ_C 和式(11)中 λ_R 都是训练时对比学习的权重，是随着训练过程动态变化的。实验发现，若初始训练时权重取正值，则模型收敛速度很慢。因此前 30 个训练批次时，权重均取 0；之后训练时权重均取 0.5，使模型能快速收敛。其次，对于未知分类阈值 ε_{un}，选择训练集上的最佳分类设置，实验中 $\varepsilon_{\text{un}}=0.82$。最后，对于再训练过程中的交叉熵阈值 ε_{re}，其取值应是训练集上小样本类分类性能最佳时的最小值，实验发现，当 $\varepsilon_{\text{re}}=1.88$ 时取得了最佳效果。

在已知恶意流量训练阶段，首先分析对比学习对检测性能（特别是小样本类）的提升效果。本节比较了本文模型、不采用对比学习的模型和采用 SimCLR 对比学习的模型的性能。如图 3 所示，在 100 次训练后，不采用对比学习的模型由于训练难度低模型已经收敛。而采用对比学习的模型，虽然训练难度高导致所需的训练次数较多，但在进一步充分训练后，模型性能还可以进一步提升。

已知类分类阶段不同设置下的归一化矩阵如图 4 所示。如图 4(a)和图 4(b)所示，加入对比学习后本文模型对部分大样本类（如良性类和 back 类）和小样本类的分类性能有明显提升。但由于加入对比学习后，模型训练难度较大，导致同样训练次数下模型可能训练不充分，极少类（如 warezclient 类）性能反而略微下降。

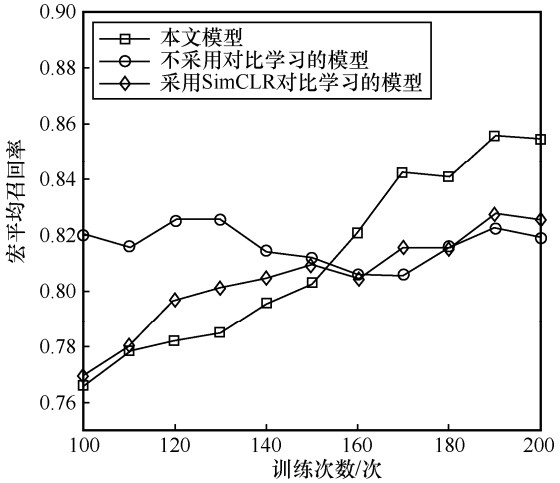


图 3 已知恶意流量训练阶段不同模型的性能变化

其次，分析再训练过程对小样本类的提升效果。如图 4(a)和 4(c)所示，再训练后模型对小样本类分类性能几乎都有明显提升。如图 4(b)和 4(d)所示，再训练过程对不采用对比学习的模型几乎没有帮助。这是因为不采用对比学习的模型对小样本恶意类分类能力较差，其输出的伪标签质量较差，对分类小样本类几乎没有帮助，反而可能降低检测性能。

在未知类分类训练阶段，本节以已知阶段本文模型为基线继续未知阶段训练过程，分别对比了本文模型和在重构阶段不采用对比学习的模型。如图 5(a)和图 5(b)所示，在重构阶段采用对比学习对未知类分类性能有一定提升，并且降低了对部分类（特别是良性类）的误判。这是因为对比学习将同类更加聚合，从而使误报减少。由于良性类在实际检测中频繁出现，这对于实际检测性能影响较大。

最后，分析模型对逃逸攻击的防御效果。实验采用 FGSM 算法对所有已知恶意类生成对抗样本测试。如图 6(a)和图 6(b)所示，相比不采样的模型，采取重采样的本文模型在同样烈度的逃逸攻击下分类性能更高，证明了重采样能有效对抗逃逸攻击。如图 6(a)和图 6(c)所示，不采用对比学习的模型也增强了对抗逃逸攻击的能力。最后，如图 6(a)和图 6(d)所示，大部分逃逸攻击即使在已知类分类阶段骗过了模型，但在未知类分类阶段逃逸攻击大多被分类为未知攻击类，也能一定程度上发现逃逸攻击。

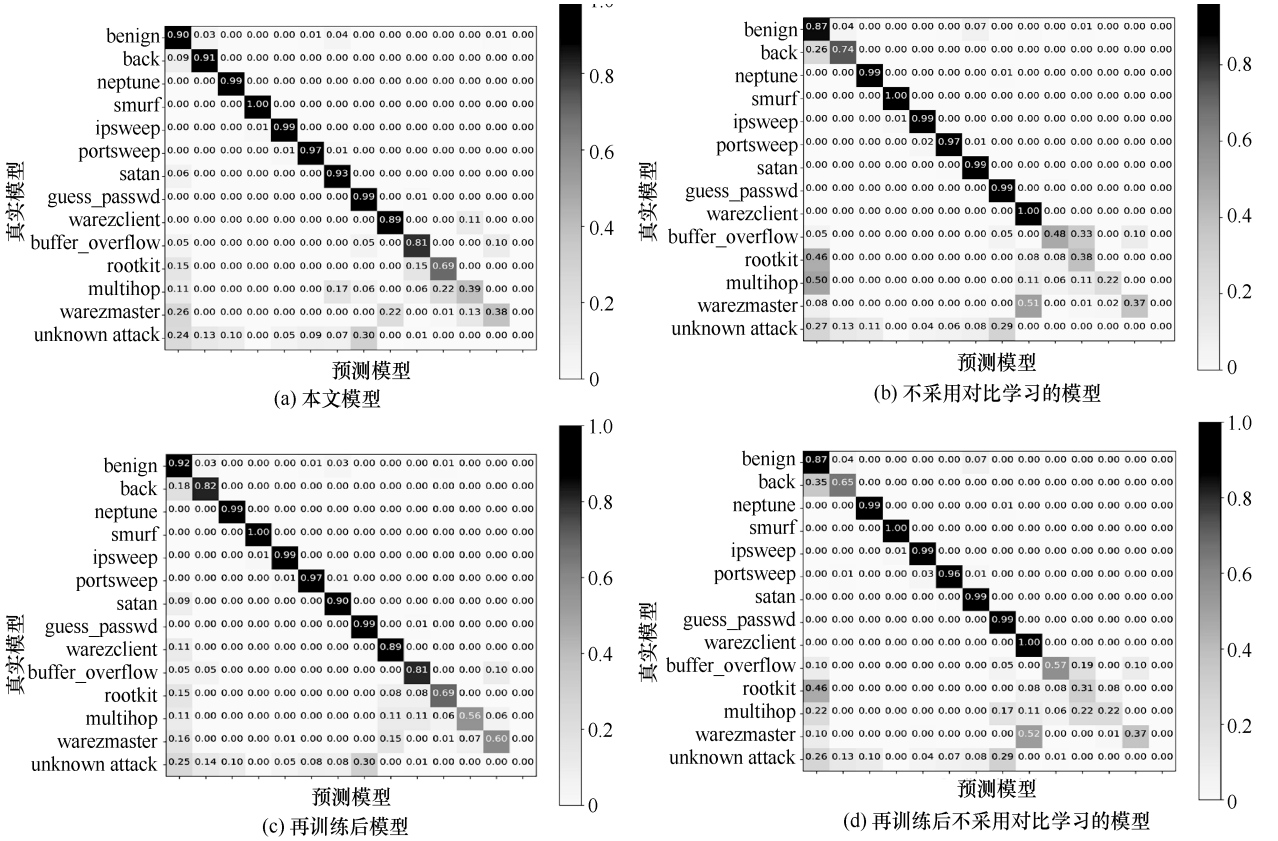


图 4　已知类分类阶段不同设置下的归一化混淆矩阵

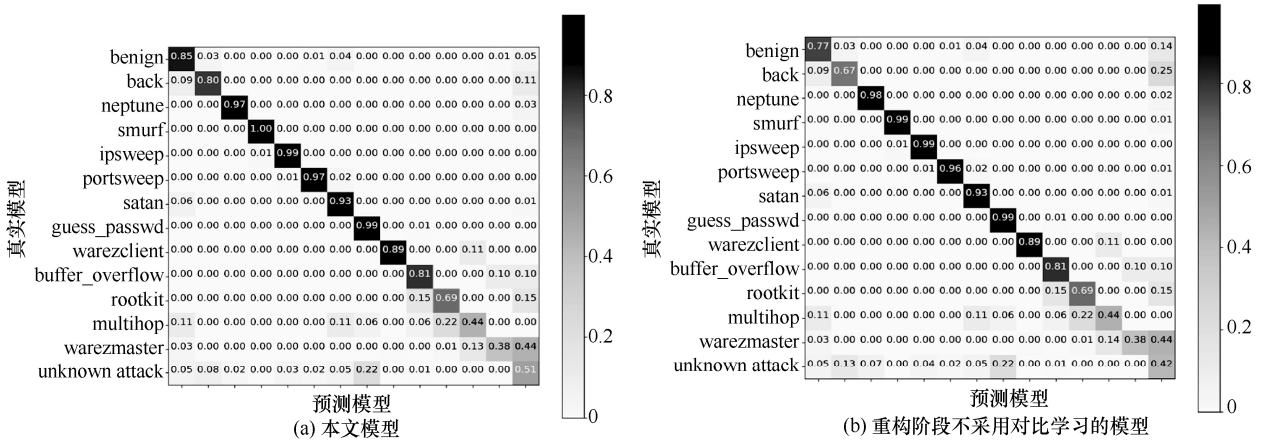
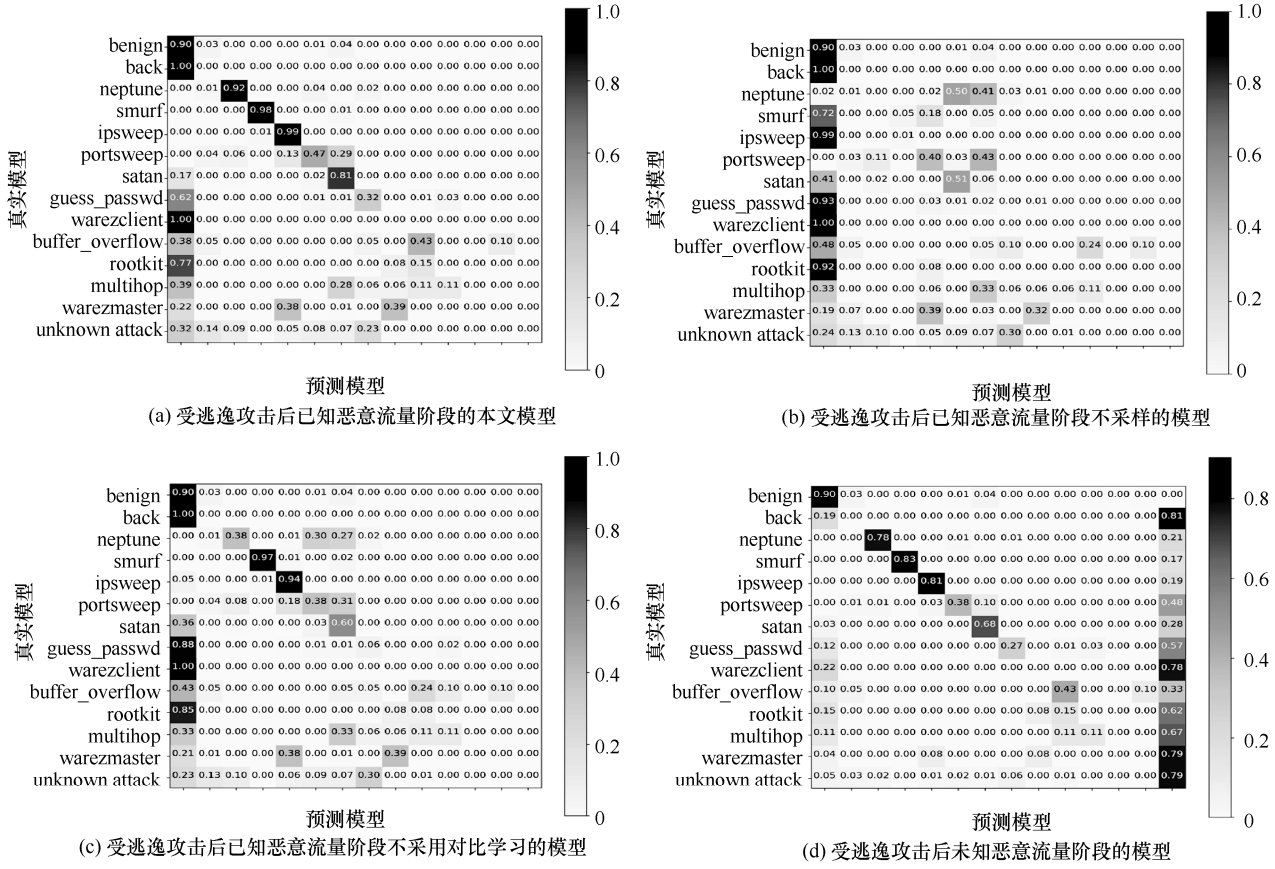


图 5　未知类分类阶段不同设置下的归一化混淆矩阵

3.4　性能分析

本节讨论本文方法与其他方法的性能对比。其中，基线模型是先进的 CVAE-EVT 模型[8]。在此实验结果与文献[8]中不同的原因主要包括以下两点：一是数据预处理方式和 MLP 网络结构不同；二是采用价标准不同，本文采用宏平均指标评。

在已知类分类阶段，分类模型都能实现对大样本已知恶意类的分类。但实际中由于网络欺骗流量或业务变化，往往存在概念漂移等问题，模型需要能快速训练并应用。因此，本节在此主要测试在相同训练时间内（本文模型在实验环境下训练 200 次的时间——约 5 min），几种方法对良性类、大样本已知类和小样本已知类的宏平均指标如表 2 所示。表 2 中，尽管随机森林方法在大样本类部分指标得到了更优结果，但不能兼顾精度与召回率。这是因为随机森林方法未采取措施平衡稀疏类，导致对稀疏（小样本）类精度提升但召回率下降，对大样本类则反之。实际中对于出现次数较少的恶意类，高召回率则代表了低漏报率，



(a) 受逃逸攻击后已知恶意流量阶段的本文模型　(b) 受逃逸攻击后已知恶意流量阶段不采样的模型

(c) 受逃逸攻击后已知恶意流量阶段不采用对比学习的模型　(d) 受逃逸攻击后未知恶意流量阶段的模型

图 6　不同设置下遭受逃逸攻击后的归一化混淆矩阵

表 2　**已知类分类阶段不同模型检测性能对比**

方法	良性类			大样本类			小样本类		
	精度	召回率	F_1 值	精度	召回率	F_1 值	精度	召回率	F_1 值
随机森林	0.817	0.954	0.880	0.829	0.780	0.705	0	0	0
MLP	0.758	0.882	0.815	0.533	0.895	0.628	0.376	0.232	0.244
CVAE[8]	0.871	0.866	0.868	0.561	0.959	0.708	0.348	0.364	0.289
本文方法	0.876	0.903	0.890	0.581	0.960	0.723	0.410	0.567	0.476

这对攻击检测而言更为重要。且随机森林对小样本类束手无策，而本文模型由于加入了对比学习以及再训练方法，在各阶段都表现出了最佳综合性能，特别是在小样本类的分类上提升明显。

在未知类分类训练阶段，相同训练时间内本文模型与先进的 CVAE-EVT 模型[8]进行了对比。图 7 给出了不同未知类设置下的宏平均 F_1 值。加入对比学习的本文模型对各类未知类的检测性能基本优于 CVAE-EVT 模型，证明了本文方法的先进性。

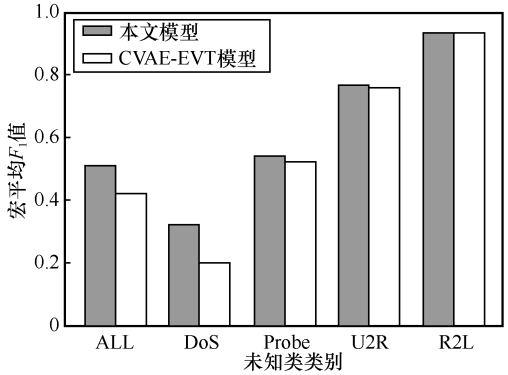


图 7　不同未知类设置下的宏平均 F_1 值

4　结束语

本文旨在设计一种应用于 NIDS 的细粒度恶意流量分类方法。在 CVAE-EVT 模型[8]的基础上，本文提出了一个能够兼顾已知类、小样本类和未知类的方法，且所设计模型需具有较强的泛化性，能够在一定程度上对抗逃逸攻击。该方法通过在各阶段引入了对比学习策略提高分类性能。并且，融合再训练、重采样、L2 正则化等技巧，使模型泛化性能进一步提高。实验结果证明了该方法的有效性和先进性。最后，由于当前逃逸攻击技术不断发展，未来拟在此基础上进一步研究在保证分类性能条件下的防御逃逸攻击流量的方法。

参考文献：

[1] SOYSAL M, SCHMIDT E G. Machine learning algorithms for accurate flow-based network traffic classification: evaluation and comparison[J]. Performance Evaluation, 2010, 67(6): 451-467.

[2] DUSI M, GRINGOLI F, SALGARELLI L. Quantifying the accuracy of the ground truth associated with Internet traffic traces[J]. Computer Networks, 2011, 55(5): 1158-1167.

[3] 陈明豪, 祝跃飞, 芦斌, 等. 基于Attention-CNN的加密流量应用类型分类[J]. 计算机科学, 2021, 48(4): 325-332.
CHEN M H, ZHU Y F, LU B, et al. Classification of application type of encrypted traffic based on attention-CNN[J]. Computer Science, 2021, 48(4): 325-332.

[4] CAMPFIELD M. The practical difference between known and unknown threats[J]. Computer Fraud & Security, 2021(5): 6-9.

[5] FRANK J. Artificial intelligence and intrusion detection: current and future directions[J]. Computers & Security, 1995, 14(1): 31.

[6] TING C, FIELD R, FISHER A, et al. Compression analytics for classification and anomaly detection within network communication[J]. IEEE Transactions on Information Forensics and Security, 2019, 14(5): 1366-1376.

[7] 曾勇, 吴正远, 董丽华, 等. 加密流量中的恶意流量分类技术[J]. 西安电子科技大学学报, 2021, 48(3): 170-187.
ZENG Y, WU Z Y, DONG L H, et al. Research on malicious traffic identification technology in encrypted traffic[J]. Journal of Xidian University, 2021, 48(3): 170-187.

[8] YANG J, CHEN X, CHEN S W, et al. Conditional variational auto-encoder and extreme value theory aided two-stage learning approach for intelligent fine-grained known/unknown intrusion detection[J]. IEEE Transactions on Information Forensics and Security, 2021, 16: 3538-3553.

[9] AKHTAR N, MIAN A. Threat of adversarial attacks on deep learning in computer vision: a survey[J]. IEEE Access, 2018, 6: 14410-14430.

[10] 韩宇, 方滨兴, 崔翔, 等. StealthyFlow: 一种对抗条件下恶意代码动态流量伪装框架[J]. 计算机学报, 2021, 44(5): 948-962.
HAN Y, FANG B X, CUI X, et al. StealthyFlow: a framework for malware dynamic traffic camouflaging in adversarial environment[J]. Chinese Journal of Computers, 2021, 44(5): 948-962.

[11] LIU J Y, ZENG Y Z, SHI J Y, et al. MalDetect: a structure of encrypted malware traffic detection[J]. Computers, Materials & Continua, 2019, 60(2): 721-739.

[12] 胡永进, 郭渊博, 马骏, 等. 基于对抗样本的网络欺骗流量生成方法[J]. 通信学报, 2020, 41(9): 59-70.
HU Y J, GUO Y B, MA J, et al. Method to generate cyber deception traffic based on adversarial sample[J]. Journal on Communications, 2020, 41(9): 59-70.

[13] DIXON L, RISTENPART T, SHRIMPTON T. Network traffic obfuscation and automated Internet censorship[J]. IEEE Security & Privacy, 2016, 14(6): 43-53.

[14] 姚忠将, 葛敬国, 张潇丹, 等. 流量混淆技术及相应分类、追踪技术研究综述[J]. 软件学报, 2018, 29(10): 3205-3222.
YAO Z J, GE J G, ZHANG X D, et al. Research review on traffic obfuscation and its corresponding identification and tracking technologies[J]. Journal of Software, 2018, 29(10): 3205-3222.

[15] HADSELL R, CHOPRA S, LECUN Y. Dimensionality reduction by learning an invariant mapping[C]//Proceedings of 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2006: 1735-1742.

[16] SOHN K, YAN X C, LEE H. Learning structured output representation using deep conditional generative models[C]//Proceedings of the 28th International Conference on Neural Information Processing Systems. Massachusetts: MIT Press, 2015: 3483-3491.

[17] HAAN L D, FERREIRA A. Extreme value theory: an introduction[M]. New York: Springer, 2006.

[18] 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. Piscataway: IEEE Press, 2016: 581-585.

[19] 潘吴斌, 程光, 郭晓军, 等. 网络加密流量分类研究综述及展望[J]. 通信学报, 2016, 37(9): 154-167.
PAN W B, CHENG G, GUO X J, et al. Review and perspective on encrypted traffic identification research[J]. Journal on Communications, 2016, 37(9): 154-167.

[20] WANG S S, YAN Q B, CHEN Z X, et al. Detecting android malware leveraging text semantics of network flows[J]. IEEE Transactions on Information Forensics and Security, 2018, 13(5): 1096-1109.

[21] YAO Z J, GE J G, WU Y L, et al. Encrypted traffic classification based on Gaussian mixture models and hidden Markov models[J]. Journal of Network and Computer Applications, 2020, 166: 102711.

[22] TAYLOR V F, SPOLAOR R, CONTI M, et al. AppScanner: automatic fingerprinting of smartphone APPs from encrypted network traffic[C]//Proceedings of 2016 IEEE European Symposium on Security and Privacy (EuroS&P). Piscataway: IEEE Press, 2016: 439-454.

[23] BAZUHAIR W, LEE W. Detecting malign encrypted network traffic using perlin noise and convolutional neural network[C]//Proceedings of 2020 10th Annual Computing and Communication Workshop and Conference (CCWC). Piscataway: IEEE Press, 2020: 200-206.

[24] SHAREVSKI F, JACHIM P, FLOREK K. To tweet or not to tweet: covertly manipulating a Twitter debate on vaccines using malware-induced misperceptions[C]//Proceedings of the 15th International Conference on Availability, Reliability and Security. Piscataway: IEEE Press, 2020: 1-12.

[25] YAO H P, FU D Y, ZHANG P Y, et al. MSML: a novel multilevel semi-supervised machine learning framework for intrusion detection system[J]. IEEE Internet of Things Journal, 2019, 6(2): 1949-1959.

[26] GENG C X, HUANG S J, CHEN S C. Recent advances in open set recognition: a survey[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021, 43(10): 3614-3631.

[27] ZHANG H, PATEL V M. Sparse representation-based open set recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(8): 1690-1696.

[28] RUDD E M, JAIN L P, SCHEIRER W J, et al. The extreme value machine[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 40(3): 762-768.

[29] BENDALE A, BOULT T E. Towards open set deep networks[C]//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2016: 1563-1572.

[30] NEAL L, OLSON M, FERN X, et al. Open set learning with counterfactual images[C]//Proceedings of the European Conference on Computer Vision (ECCV). Berlin: Springer, 2018: 613-628.

[31] GENG C X, CHEN S C. Collective decision for open set recognition[J]. IEEE Transactions on Knowledge and Data Engineering, 2022, 34(1): 192-204.

[32] CRUZ S, COLEMAN C, RUDD E M, et al. Open set intrusion recognition for fine-grained attack categorization[C]//Proceedings of 2017 IEEE International Symposium on Technologies for Homeland Security. Piscataway: IEEE Press, 2017: 1-6.

[33] HENRYDOSS J, CRUZ S, RUDD E M, et al. Incremental open set intrusion recognition using extreme value machine[C]//Proceedings of 2017 16th IEEE International Conference on Machine Learning and Applications. Piscataway: IEEE Press, 2017: 1089-1093.

[34] SZEGEDY C, ZAREMBA W, SUTSKEVER I, et al. Intriguing properties of neural networks[J]. arXiv Preprint, arXiv: 1312.6199, 2013.

[35] GOODFELLOW I J, SHLENS J, SZEGEDY C. Explaining and harnessing adversarial examples[J]. arXiv Preprint, arXiv: 1412.6572, 2014.

[36] MOPURI K R, GARG U, BABU R V. Fast feature fool: a data independent approach to universal adversarial perturbations[J]. arXiv Preprint, arXiv: 1707.05572, 2017.

[37] LIN Z, SHI Y, XUE Z. Idsgan: generative adversarial networks for attack generation against intrusion detection[J]. arXiv Preprint, arXiv:1809.02077, 2018.

[38] LI J, ZHOU L, LI H X, et al. Dynamic traffic feature camouflaging via generative adversarial networks[C]//Proceedings of 2019 IEEE Conference on Communications and Network Security. Piscataway: IEEE Press, 2019: 268-276.

[39] GRILL J B, STRUB F, ALTCHÉ F, et al. Bootstrap your own latent-a new approach to self-supervised learning[J]. arXiv Preprint, arXiv: 2006.07733, 2020.

[40] SOHN K. Improved deep metric learning with multi-class N-pair loss objective[C]//Proceedings of the 30th International Conference on Neural Information Processing Systems. Massachusetts: MIT Press, 2016: 1857-1865.

[41] CHEN T, KORNBLITH S, NOROUZI M, et al. A simple frame-work for contrastive learning of visual representations[J]. arXiv Preprint, arXiv: 2002.05709, 2020.

[42] HE K M, FAN H Q, WU Y X, et al. Momentum contrast for unsupervised visual representation learning[C]//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2020: 9726-9735.

[43] HU Q J, WANG X, HU W, et al. AdCo: adversarial contrast for efficient learning of unsupervised representations from self-trained negative adversaries[C]//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2021: 1074-1083.

[44] HO C H, NVASCONCELOS N. Contrastive learning with adversarial examples[J]. arXiv Preprint, arXiv: 2020.12050, 2020.

[45] LI M Z, LIN X X, CHEN X Y, et al. Keywords and instances: a hierchical contrastive learning framework unifying hybrid granularities for text generation[J]. arXiv Preprint, arXiv:2205.13346, 2022.

[46] BUKCHIN G, SCHWARTZ E, SAENKO K, et al. Fine-grained angular contrastive learning with coarse labels[C]//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2021: 8726-8736.

[47] HABIBI L A H, DRAPER-GIL G, MAMUN M S I, et al. Characterization of tor traffic using time based features[C]//Proceedings of the 3rd International Conference on Information Systems Security and Privacy. [S.l.]: SCITEPRESS - Science and Technology Publications, 2017: 253-262.

[48] HIGGINS I, MATTHEY L, PAL A, et al. Beta-vae: learning basic visual concepts with a constrained variational framework[C]//ICLR 2017 Conference Homepage. [S.l.:s.n.], 2017: 1-8.

[49] OORD A V D, LI Y Z, VINYALS O. Representation learning with contrastive predictive coding[J]. arXiv Preprint, arXiv: 1807.03748, 2018.

[50] HENDRYCKS D, GIMPEL K. Early methods for detecting adversarial images[J]. arXiv Preprint, arXiv:1608.00530, 2016.

[51] FEINMAN R, CURTIN R R, SHINTRE S, et al. Detecting adversarial samples from artifacts[J]. arXiv Preprint, arXiv:1703.00410, 2017.

[52] TANAY T, GRIFFIN L D. A new angle on L2 regularization[J]. arXiv Preprint, arXiv:1806.11186, 2018.

[53] TRAMÈR F, KURAKIN A, PAPERNOT N, et al. Ensemble adver-

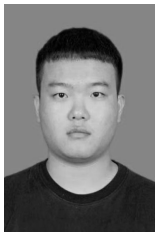
sarial training: Attacks and defenses[J]. arXiv Preprint, arXiv: 1705.07204, 2017.

[54] TAVALLAEE M, BAGHERI E, LU W, et al. A detailed analysis of the KDD CUP 99 data set[C]//Proceedings of 2009 IEEE Symposium on Computational Intelligence for Security and Defense Applications. Piscataway: IEEE Press, 2009: 1-6.

[作者简介]

王一丰（1994− ），男，江苏泰兴人，信息工程大学博士生，主要研究方向为零样本学习、网络安全和入侵检测等。

郭渊博（1975− ），男，陕西周至人，博士，信息工程大学教授、博士生导师，主要研究方向为大数据安全、态势感知等。



陈庆礼（1998− ），男，河南新乡人，信息工程大学硕士生，主要研究方向为人工智能安全等。

方晨（1993− ），男，安徽宿松人，博士，信息工程大学讲师，主要研究方向为机器学习、隐私安全等。

林韧昊（1993− ），男，河南郑州人，郑州大学博士生，主要研究方向为深度学习、鲁棒性验证和网络安全等。