DOI:10.11918/202407062

基于对比学习和关系蒸馏的增量式入侵检测方法

金志刚,刘 凯,陈旭阳,武晓栋

(天津大学 电气自动化与信息工程学院,天津 300072)

摘 要:为解决入侵检测系统在学习新类攻击样本时存在的灾难性遗忘、分类偏向新类、旧类泛化能力弱等问题,提出基于对比学习和关系蒸馏的增量式入侵检测方法(incremental intrusion detection method based on contrastive learning and relational distillation,CLRD-IIDM)。首先,提出对比学习增强方法,构建具有特征提取层、投影分支和分类分支的双分支网络结构,并提出旧类补偿对比损失与之配合以增强对旧类样本的记忆;其次,提出新旧类关系蒸馏方法,提取旧模型中的新旧类相似关系并构造类关系标签,利用这些标签对新模型施加正则化约束,缓解模型对新类的分类偏向问题;最后,设计均衡偏差示例存储器,存储能够有效且全面表征旧类特征的样本以保持模型对旧类的泛化能力。在 UNSW-NB15 数据集上的实验结果表明,应用 CLRD-IIDM 的多个神经网络模型的最终准确率优于基线模型 2.11% ~9.78%;对比实验和消融实验结果验证了本文方法的有效性和泛化性。CLRD-IIDM 能较好地适应复杂多变的现实网络环境,具有一定的应用价值。

关键词:入侵检测系统;灾难性遗忘;增量学习;对比学习;知识蒸馏

中图分类号:TP393.08 **文献标志码**:A **文章编号**:0367-6234(2026)09-0060-10

An incremental intrusion detection method based on contrastive learning and relational distillation

Jin Zhigang, Liu Kai, Chen Xuyang, Wu Xiaodong

(School of Electrical and Information Engineering, Tianjin University, Tianjin 300072, China)

Abstract: To solve the problems of catastrophic forgetting, classification bias towards new classes, and weak generalization ability of old classes in intrusion detection systems when learning new attack samples, an incremental intrusion detection method based on contrastive learning and relational distillation (CLRD-IIDM) is proposed. Firstly, a contrastive learning enhancement method is introduced, which constructs a dual branch network structure with the feature extraction layer, projection branch, and classification branch. The old class compensation contrastive loss is employed to cooperate with it to enhance the memory of old class samples; Secondly, the distillation method for the relationship between old and new classes is proposed, which extracts the similarity relationship between old and new classes in the old model and constructs class relationship labels. These labels are used to apply regularization constraints to the new model to alleviate the problem of classification bias towards new classes; Finally, the balanced bias example memory is designed to store samples that can effectively and comprehensively characterize the features of the old class, which maintains the model's generalization ability to the old class. The experimental results on the UNSW-NB15 dataset show that the final accuracy of multiple neural network models using CLRD-IIDM is 2.11% to 9.78% better than the baseline model. Comparative experiments and ablation experiments have demonstrated the effectiveness and generalization of the proposed method. CLRD-IIDM can adapt well to complex and ever-changing real-world network environments, and has certain practical value.

Keywords: intrusion detection system; catastrophic forgetting; incremental learning; contrastive learning; knowledge distillation

互联网、物联网等各类网络技术的大规模应用使得网络环境变得日益多样化和复杂化,导致网络空间安全面临严峻挑战。入侵检测系统(intrusion detection system,IDS)作为一种重要的安全防御手段,已成为研究的焦点[1]。

基于深度学习的 IDS 能够充分提取网络流量特

收稿日期:2024-07-22;**录用日期**:2024-08-28;**网络首发日期**:2025-06-12
网络首发地址:https://link.cnki.net/urlid/23.1235.t.20250611.1804.006
基金项目:国家自然科学基金(52171337)
作者简介:金志刚(1972—),男,教授,博士生导师
通信作者:金志刚,zgjin@tju.edu.cn

征，且检测性能优秀，被广泛用于 IDS 中网络攻击检测。IDS 依赖训练样本获得对网络攻击的识别能力，在不完备数据集下训练难以适应网络流量动态变化、新类别攻击不断涌现[2-3]的现实网络环境。同时，利用新类数据对 IDS 进行微调会遭遇深度学习固有的灾难性遗忘问题[4]，即 IDS 在拟合新类数据的过程中会丧失对旧类的识别能力。联合新类数据和旧类数据重新训练 IDS 是一种有效的方法，然而，旧类数据可能会因存储成本过大和隐私问题而不可用。此外，随着新类别攻击的出现，反复重新训练 IDS 的高成本和低效率也使得这一方法难以在实际中应用。

增量学习方法（incremental learning，IL）允许 IDS 在学习新类数据的同时保留对旧类的识别能力[5]。国内外研究人员对基于增量学习的 IDS 展开了一系列研究。王一丰等[6]提出一种细粒度恶意流量识别方法，该方法利用变分自编码器存储原有类知识，通过旧类样本重构和知识蒸馏方法实现对原有类检测能力的保持。Amalapuram 等[7]提出基于增强内存回放的增量学习方法并应用于入侵检测，该方法提出一种扩展的类平衡水库抽样算法优化旧样本存储，并提出一种参数近似扰动辅助算法优化训练过程。Du 等[8]提出一种基于分支融合策略的少样本增量式 IDS，该系统仅使用少量样本增量学习新的网络攻击，并在增量学习过程中微调系统参数以避免灾难性遗忘问题。综上，增量学习的相关方法主要分为两类：一类是基于正则化[9-10]的方法，即引入正则项限制模型参数的优化方向，保障模型学习到的旧类知识不会被新类知识破坏；一类是基于回放[11-13]的方法，即设置一个内存受限的示例存储器存储旧类样本，以极小的内存代价帮助模型保留旧类知识。在入侵检测任务中，基于正则化的方法由于不访问旧类数据，无法有效应对不断出现新类样本带来的数据特征分布变化问题，对旧类泛化能力较弱，因此常与基于回放的方法结合使用。而基于回放的方法仅允许模型访问少量旧类样本，在训练时，新旧类样本数据严重失衡，模型倾向于适应新类样本，分类偏向新类[14]。因此，缓解模型的分类偏向，增强模型对旧类的泛化性成为研究的关键。

基于此，本文提出基于对比学习和关系蒸馏的增量式入侵检测方法（incremental intrusion detection method based on contrastive learning and relational distillation，CLRD-IIDM）。提出对比学习增强方法，同时优化模型的特征提取能力和分类性能，并构造旧类补偿对比损失与之配合以缓解灾难性遗忘问题。提出新旧类关系蒸馏方法，缓解分类偏向问题。设计均衡偏差示例存储器，确保所存储旧类样本的特征表征能力。在多个神经网络模型上应用所提方法，并与基线模型对比以验证所提方法的有效性和泛化性。此外，通过消融实验探究所提方法中各部分对于增量式 IDS 识别性能提升的具体效用。

1　对比学习增强方法

对比学习[15-17]通过构建正负样本对，并在特征空间对齐语义相同的正样本对，分离语义不同的负样本对的方式增强模型的表征能力，提升模型的性能。因此，本文提出对比学习增强方法，借助对比学习的特征对齐能力学习样本的本质特征，增强模型的特征提取能力以缓解模型的灾难性遗忘问题。对比学习增强方法如图 1 所示。研究[17]表明，利用投影头将数据特征投影至低维潜在特征空间可以过滤原始特征中可能存在的噪声，并进一步提取数据的可泛化性特征，从而增强数据特征的可区分性。因此，在对比学习增强方法模型中插入一个投影头作为投影分支，构成由特征提取层、投影分支和分类分支组成的双分支网络结构。在投影分支中，应用对比损失优化模型的特征提取能力；在分类分支中，应用交叉熵损失优化模型的分类性能。随后，提出旧类补偿对比损失，依据分类分支对各类别的学习效果以及新旧类数量差异调整投影分支对不同类别数据特征的学习权重，对数量较少的旧类进行权重补偿，从而使模型学习到各类更具区分度的数据特征。反过来，模型特征提取能力的增强则有助于分类分支进行更好地分类，从而提升 CLRD-IIDM 的性能。

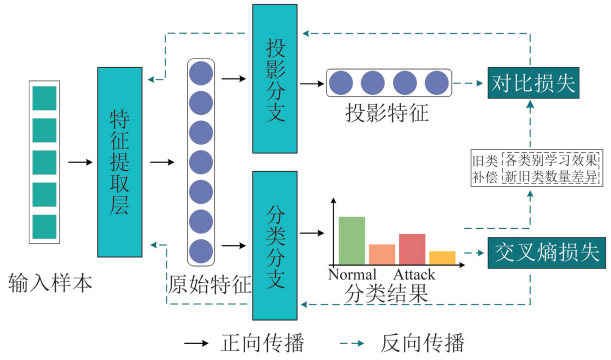


图 1　对比学习增强方法

Fig. 1　Contrastive learning enhancement method

1.1　双分支网络结构

双分支网络结构的具体组成如下：

1）特征提取层 $f_{\mathrm{FE}}(\cdot)$：将网络流量输入 $f_{\mathrm{FE}}(\cdot)$，提取数据原始特征。对于给定的网络流量样本 x，特征提取层输出的原始特征可以表示为

$$h = f_{\mathrm{FE}}(x) \tag{1}$$

2)投影分支 $g_{\mathrm{Pr}}(\cdot)$:投影分支由两层全连接层组成。特征提取层提取到的数据特征被投影分支映射到应用旧类补偿对比损失的潜在特征空间。对于给定的数据特征,投影分支输出的投影特征可以表示为

$$z = g_{\mathrm{Pr}}(h) = W^{(2)}\sigma(W^{(1)}h) \tag{2}$$

式中:$W^{(1)}$、$W^{(2)}$ 为两层全连接层的权重;$\sigma(\cdot)$ 为 ReLU 激活函数,用于将投影分支转换成非线性结构以增强其学习与表征能力。

3)分类分支 $f_{\mathrm{CLS}}(\cdot)$:基于特征提取层的输出,分类分支采用全连接神经网络对网络流量的数据类别进行分类,检测当前网络流量是否属于网络攻击。对于给定的网络流量样本 x,分类分支输出可以表示为

$$y' = f_{\mathrm{CLS}}(h) = f_{\mathrm{CLS}}(f_{\mathrm{FE}}(x)) \tag{3}$$

为便于下文清晰表述,用 $\boldsymbol{\Psi}$ 表示由特征提取层和投影分支组成的投影网络,用 $\boldsymbol{\Theta}$ 表示由特征提取层和分类分支组成的分类模型,用 Λ 表示由特征提取层、投影分支和分类分支组成的完整模型。

1.2 旧类补偿对比损失

在 IDS 的增量学习任务中,新旧类别数据数量失衡问题导致模型倾向于学习新类,旧类则发生不同程度的灾难性遗忘,而传统对比损失无法针对这一问题对各类别样本做出对应调整[17]。针对此局限性,本文提出旧类补偿对比损失优化模型。在模型对齐正样本对特征、分离负样本对特征时,依据模型对样本拟合程度的不同,对各类别样本进行重加权,数量较少、易被遗忘的旧类样本被重点关注以保持模型对旧类样本的记忆,进而缓解模型的灾难性遗忘问题。

具体而言,入侵检测模型当前处理任务 T^n 下的一个 mini-batch 数据 B^n 表示为 $\{X_b^n, Y_b^n\} = \{x_i^n, y_i^n\}_{i=1}^b \subset D^n \cup D^{\mathrm{M},n}$,其中 D^n 为当前任务的新类数据,$D^{\mathrm{M},n}$ 为示例存储器中存储的旧类样本,y_i^n 为对应样本的标签,b 为 B^n 的 batch-size。此 mini-batch 的传统对比损失 L_{CON} 为:

$$L_{\mathrm{CON}} = \frac{1}{b}\sum_{i\in I} L_{\mathrm{CON},i} \tag{4}$$

$$L_{\mathrm{CON},i} = \frac{-1}{|C(i)|}\sum_{p\in C(i)} \log\frac{\exp(z_i\cdot z_c/\tau)}{\sum\limits_{a\in A(i)}\exp(z_i\cdot z_a/\tau)} \tag{5}$$

式中:$i \in I = \{1,2,\cdots,b\}$ 为 B^n 中各样本的索引,$A(i) = \{a\in I: a\neq i\}$ 为 B^n 中除样本 x_i 外其他样本的索引,$C(i) = \{c\in A(i): y_c = y_i\}$ 为 B^n 中与样本 x_i 有相同标签的样本的索引,$|C(i)|$ 为 $C(i)$ 的基数,z_i、z_c、z_a 分别为 i、c、a 对应类型的样本在潜在特征空间中的数据特征,τ 为温度超参数。

由式(4)和式(5)可以看出,面对 IDS 增量学习任务的灾难性遗忘问题时,传统对比损失缺乏对样本学习效果差异和新旧类数据量差异的考量。因此,提出旧类补偿对比损失应对此问题。具体而言,模型对样本 (x_i^n, y_i^n) 的拟合程度 h_i^n 定义为

$$h_i^n = f_{\mathrm{CE}}(P^n(x_i^n, \boldsymbol{\Theta}^n), y_i^n) \tag{6}$$

式中:$P^n(x_i^n, \boldsymbol{\Theta}^n)$ 为样本 (x_i^n, y_i^n) 经 $\boldsymbol{\Theta}^n$ 分类得到的概率分布,$f_{\mathrm{CE}}(\cdot)$ 为交叉熵函数。

根据式(6)可知,模型对样本的拟合程度越差,h_i^n 越大,表明模型还没有很好地学习到该样本或者已经遗忘该样本的知识,需要加强对该样本的学习;模型对样本的拟合程度越好,h_i^n 越小,表明模型对该样本的知识掌握得越好,则需要放缓对该样本的学习。

进一步,新旧类的整体拟合程度定义为:

$$H_{\mathrm{new}}^n = \sum_{i=1}^{b} h_i^n \kappa(y_i \in Y_{\mathrm{new}}) \tag{7}$$

$$H_{\mathrm{old}}^n = \sum_{i=1}^{b} h_i^n \kappa(y_i \in Y_{\mathrm{old}}) \tag{8}$$

式中:$\kappa(\cdot)$ 为条件判断函数,当变量为真时,$\kappa(\cdot)=1$,为假时则 $\kappa(\cdot)=0$;Y_{new} 和 Y_{old} 分别为新类标签和旧类标签。

利用式(7)和式(8)计算新旧类的整体学习效果。

基于此,对于一个 mini-batch 数据 B^n 的旧类补偿对比损失 $L_{\mathrm{OCC\text{-}CON}}$ 定义为

$$L_{\mathrm{OCC\text{-}CON}} = \sum_{i\in I} w_i^n L_{\mathrm{CON},i} = \sum_{i\in I}\frac{h_i^n}{H_i^n} L_{\mathrm{CON},i} \tag{9}$$

式中 $H_i^n = H_{\mathrm{new}}^n \kappa(y_i \in Y_{\mathrm{new}}) + H_{\mathrm{old}}^n \kappa(y_i \in Y_{\mathrm{old}})$。

通过引入权重 $\dfrac{h_i^n}{H_i^n}$,$L_{\mathrm{OCC\text{-}CON}}$ 将依据旧类数量和旧类整体学习效果对数量较少的旧类进行学习补偿,增加其在对比损失中的权重占比。而对于新旧类内部的每个类别,$L_{\mathrm{OCC\text{-}CON}}$ 根据每个类别样本的学习效果平衡学习速度,重点处理难以学习或易被遗忘的样本,缓解模型在训练过程中的灾难性遗忘问题。

2 新旧类关系蒸馏方法

旧分类模型 $\boldsymbol{\Theta}^{n-1}$ 中保存着旧类数据知识。当采用旧模型 $\boldsymbol{\Theta}^{n-1}$ 预测新类数据类别时,$\boldsymbol{\Theta}^{n-1}$ 输出的概率分布可以看作新旧类之间的相似关系。本文利用这种类间相似关系对新类学习施加正则化约束,缓解模型的分类偏向问题。具体而言,在当前任务 T^n 下,利用上一个任务 T^{n-1} 的旧模型 $\boldsymbol{\Theta}^{n-1}$ 对一个

mini-batch 数据 $\{X_b^n, Y_b^n\} = \{x_i^n, y_i^n\}_{i=1}^b \subset D^n \cup D^{M,n}$ 进行分类，获得 Θ^{n-1} 对该 mini-batch 数据预测的概率分布 $P^{n-1}(x_i^n, \Theta^{n-1})$，该概率分布表示由 Θ^{n-1} 得出的新类样本与各旧类之间的相似程度。进一步，每个新类与各旧类之间的类关系原型计算公式定义为

$$\tilde{P}_k^n = \frac{1}{\phi_k}\sum_{i=1}^{b} P^{n-1}(x_i^n, \Theta^{n-1})\kappa(y_i^n = k) \quad (10)$$

$$\phi_k = \sum_{i=1}^{b} \kappa(y_i^n = k) \quad (11)$$

通过计算每个新类与各旧类之间的类关系原型来消除可能存在的异常样本引入的噪声。

接下来，将 X_b^n 的标签 Y_b^n 独热编码，并将其前 $C\cdot(n-1)$ 个维度按照对应类别替换为 $\tilde{P}_k^n$（C 为每个增量任务中的样本类别数），获得蕴含新旧类相似关系的类关系标签 $Y_{b,r}^n$。最后，利用 $Y_{b,r}^n$ 对新模型 Θ^n 进行类关系知识蒸馏。类关系知识蒸馏损失表达式为

$$L_{RD} = D_{KL}(P^n(x_i^n, \Theta^n), Y_{b,r}^n) \quad (12)$$

式中 $D_{KL}(\cdot)$ 为 KL 散度。

类关系标签对模型新类学习施加正则化约束，使得模型权重更新方向不再由新类过度主导，增强了模型对旧类知识的学习，缓解了模型对新类的分类偏向问题。具体而言，利用类关系标签对模型进行知识蒸馏这一操作使得模型在拟合新类时不仅学习新类独热编码标签中的分类信息，更新与新类相关的权重，还从类关系标签中学习新类与各旧类之间的相似关系，更新与旧类相关的权重。

此外，新旧类关系蒸馏方法作用于分类分支，与常规的交叉熵损失 L_{CE} 共同优化模型参数。因此，分类分支所用的损失函数为

$$L_{CLS} = \alpha_{CE}L_{CE} + \alpha_{RD}L_{RD} \quad (13)$$

式中 α_{CE} 和 α_{RD} 为超参数，用于平衡新旧类关系蒸馏方法和交叉熵损失的权重。

3　均衡偏差示例存储器

在现实应用中，增量式 IDS 受限于示例存储器的内存大小，无法存储所有已学习的旧类样本，应存储对于保持旧类知识尤为有益的旧类样本，因此，本文提出均衡偏差示例存储器。该存储器利用均衡偏差示例选取方法（bias equilibrium-based sample selection method，BESS）选取有效且全面表征旧类特征的样本并进行存储。具体而言，对比学习增强方法会在潜在特征空间聚集同类样本，分离异类样本。而靠近特征中心的样本具备该类样本代表性特征，能帮助模型巩固已学习到的该类代表知识，避免特征漂移问题[18]；偏离特征中心的样本反映了该类样本中更复杂多样的情况，能帮助模型学习样本中更广泛的旧类知识[19]。存储器通过存储靠近和偏离特征中心的样本来帮助模型更好地回忆旧类知识，进而提升模型对旧类的泛化能力。

对于标签为 y、数量为 $|X|$ 的一类样本 X，其中的一条样本 x_i 经投影网络 Ψ 投影得到的特征为 $Z(x_i, \Psi)$，该样本距特征中心的偏差距离计算式为：

$$d_i = \| Z(x_i, \Psi) - \bar{Z} \|_2 \quad (14)$$

$$\bar{Z} = \frac{\sum_{i=1}^{|X|} Z(x_i, \Psi)}{|X|} \quad (15)$$

式中 $\|\cdot\|_2$ 为 L_2 范数。

在每个增量学习任务结束时，依类别计算所有样本的偏差距离。随后，采用等偏差距离间距采样的方式均衡选取新类中靠近特征中心的样本和偏离特征中心的样本进行存储，以使存储样本对旧类的特征表征更加全面。同时，依据现实数据分布更新示例存储器中各旧类的样本数量。示例存储器的内存不断随新类存储而增加，这不符合实际应用，因此，均衡偏差示例存储器的内存大小固定不变。具体而言，假设示例存储器内存为 m，每个增量任务中的样本类别数为 C，增量任务 T^t 中的各类样本为 $D_c^t \in \{D_c^t\}_{c=1}^C$，则增量任务 T^n 完成后，示例存储器中各旧类样本 $D_c^{M,n} \in \{D_c^{M,n}\}_{c=1}^{C\cdot n}$ 的数量可表示为

$$|D_c^{M,n}| = \frac{|D_c^t| \cdot m}{\sum_{t=1}^{n}\sum_{c=1}^{C}|D_c^t|} \quad (16)$$

总之，在当前任务训练结束后，利用 BESS 更新存储器中已存储的旧类样本，并选取新类样本存入存储器。BESS 伪代码和均衡偏差示例存储器的更新流程伪代码如下：

算法 1：BESS

输入：增量任务 T^n 中的投影网络 Ψ^n，等待被采样的新类或旧类样本 D，采样数量 S

输出：在样本 D 中选取均衡样本 D'

$D' = \{\}$

利用式(14)计算 D 中各样本与类中心的偏差距离 d，其中最小偏差距离为 d_{min}，最大偏差距离为 d_{max}

计算 D 中各类样本的类内最大偏差距离差：$d' = d_{max} - d_{min}$

计算采样间距：$\delta = d'/S$

For s **in** S **do**//等间距采样

　　计算待选取样本的偏差距离：$d_{ch} = d_{min} + \delta\cdot s$

　　$D' += x$　//x 为偏差距离最接近 d_{ch} 的样本

End for

Return 均衡样本 D'

算法 2:均衡偏差示例存储器更新流程

输入:内存大小为 m 的示例存储器 M^n,增量任务 T^n 的新类样本 D_c^n,示例存储器中的旧类样本 $D_c^{\mathrm{M},n-1} \in \{D_c^{\mathrm{M},n-1}\}_{c=1}^{C\cdot(n-1)}$

输出:更新后的示例存储器 M^{n+1}

If 增量任务 T^n 结束 **then**

利用式(16)计算示例存储器中应存储的各类样本数量 $|D_c^{\mathrm{M},n}|$

For c **in** $C\cdot n$ **do**

if $c \leqslant C\cdot(n-1)$ **then** //更新示例存储器中各类样本的数量

利用算法 1 选取多样性样本,存入 M^{n+1}

End for

Return 更新后的示例存储器 M^{n+1}

4 CLRD-IIDM 训练流程

CLRD-IIDM 的整体训练流程如图 2 所示。当新类攻击样本到来时,模型开始进行增量训练。在增量任务 T^n 中,模型的训练样本由新类攻击样本和示例存储器中存储的旧类样本组成。将训练样本送入模型,利用对比学习增强方法和新旧类关系蒸馏方法对模型进行增量训练。当训练结束时,利用 BESS 选取用于存储的新旧类样本并更新均衡偏差示例存储器以保持模型对旧类的泛化能力。

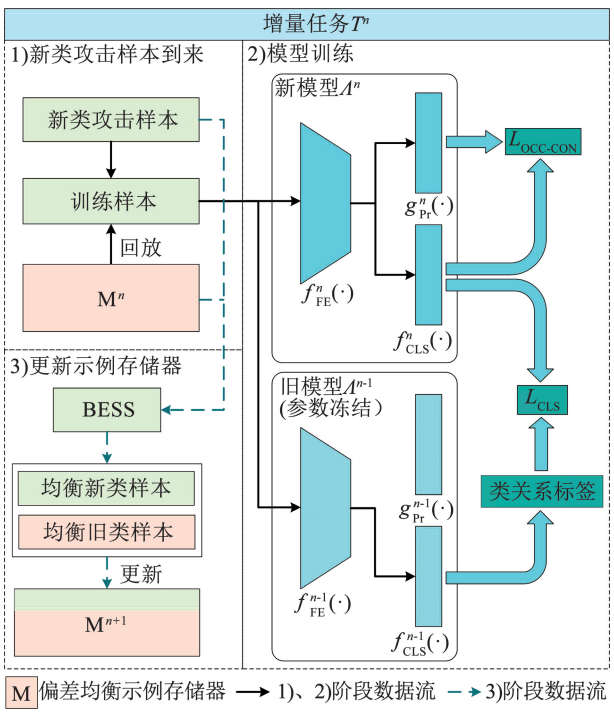


图 2 CLRD-IIDM 整体训练流程

Fig. 2 Overall training process of CLRD-IIDM

CLRD-IIDM 训练流程伪代码如下:

算法 3:CLRD-IIDM 训练流程

输入:新类攻击样本 D^n,示例存储器中旧类样本 $D^{\mathrm{M},n}$

输出:增量学习模型 Λ^n

Repeat

For mini-batch $\{X_b^n, Y_b^n\}$ **in** $D^n \cup D^{\mathrm{M},n}$ **do**

计算 $\{X_b^n, Y_b^n\}$ 在 Θ^n 上的样本拟合程度 $\{H_b^n\}$ 和新旧类整体拟合程度 H_{new}^n、H_{old}^n

根据 Θ^n 对 $\{X_b^n, Y_b^n\}$ 的拟合程度对 L_{CON} 进行旧类补偿

将 $\{X_b^n, Y_b^n\}$ 输入 Θ^{n-1} 构造类关系标签 $Y_{b,\mathrm{r}}^n$

计算 $\{X_b^n, Y_b^n\}$ 在 Θ^n 上的输出与 $Y_{b,\mathrm{r}}^n$ 之间的类关系蒸馏损失 L_{RD}

计算 $\{X_b^n, Y_b^n\}$ 在 Θ^n 上的交叉熵损失 L_{CE}

最小化投影分支上 $L_{\mathrm{OCC\text{-}CON}}$ 以更新投影网络 Ψ^n 参数

最小化分类分支上总损失 L_{CLS} 以更新分类模型 Θ^n 参数

Until 当前任务训练结束

根据算法 2 更新示例存储器

Return 增量学习模型 Λ^n

5 实验结果与分析

5.1 数据集

为了评估 CLRD-IIDM 的性能,在 UNSW-NB15 数据集[20]上开展实验。UNSW-NB15 数据集是由澳大利亚网络安全中心收集整理的,已被广泛应用于入侵检测研究之中。其共包含 10 种数据类型,即 1 种良性数据和 9 种攻击类别。实验中舍弃数据集中样本过少的 Worms 类。针对本文增量学习任务,UNSW-NB15 数据集具体划分见表 1。

表 1 UNSW-NB15 数据集具体划分

Tab. 1 Specific divisions of UNSW-NB15 dataset

任务阶段	类别	训练集样本数	测试集样本数
任务 1	Normal	56 000	36 988
	Fuzzers	18 184	6 062
	Backdoor	1 746	583
任务 2	DoS	12 264	4 089
	Exploits	33 393	11 132
	Analysis	2 000	677
任务 3	Generic	40 000	18 871
	Reconnaissance	10 491	3 496
	Shellcode	1 133	378

5.2 评价指标

采用准确率 A、精确率 P、召回率 R 作为评价指标。A 用于评价模型的整体分类性能,P 和 R 用于评价模型在识别攻击时的准确性和全面性。各评价指标的计算式如下:

$$A=\frac{T_P+T_N}{T_P+T_N+F_P+F_N} \tag{17}$$

$$P=\frac{T_P}{T_P+F_P} \tag{18}$$

$$R=\frac{T_P}{T_P+F_N} \tag{19}$$

式中：T_P为模型正确识别为攻击样本的样本数量，F_P为模型错误将正常样本识别为攻击样本的样本数量，T_N为模型正确识别为正常样本的样本数量，F_N为模型错误将攻击样本识别为正常样本的样本数量。

5.3　实验环境及参数设置

实验的硬件环境与软件环境见表 2。

表 2　实验环境

Tab. 2　Experimental environment

实验环境	设备/软件	设备型号/软件版本
硬件环境	CPU	12th Gen Intel(R) Core(TM) i5-12400 2.50 GHz
	GPU	NVIDIA GeForce RTX 3070
	机带 RAM	32 GB(3 200 MHz)
软件环境	操作系统	Windows 11
	语言	Python 3.9
	框架	PyTorch 1.13.1

实验的主要参数设置：τ 设为 0.07，存储器内存设为本文训练集数量的 2%。此外，为了便于分析实验结果，超参数 α_{CE} 和 α_{RD} 均设为 0.5。

5.4　不同模型效果对比实验

为验证本文所提方法的有效性和泛化性，以 DNN[21]、CNN[22]、RNN[23]、CNN-GRU[24] 模型作为所提方法的基础模型，并与未使用本文方法的最佳性能模型（基线）进行对比，实验结果见表 3。

由表 3 可以看出，本文方法在 CNN-GRU 上取得最佳性能，在不包含旧类数据的任务 1 中，CNN-GRU 的准确率比基线模型高 5.37%，说明对比学习增强方法提升了模型提取数据本质特征的能力，增强了模型的整体性能。此外，CNN-GRU 模型在精确率和召回率指标上的优势最为突出。其中，在任务 3 中，精确率和召回率分别比基线模型高 17.25% 和 10.93%，这意味着 CNN-GRU 模型能够更精准、更全面地识别各种攻击行为。原因是基线模型在旧类上的精确率和召回率严重衰减，有的甚至降为 0，即在该类上出现了严重的灾难性遗忘问题。而 CNN-GRU 模型在旧类补偿对比损失和新旧类关系蒸馏方法的作用下，保持了对旧类知识的记忆，对旧类的识别能力强于基线模型。对于 RNN 模型，其在任务 3 中的精确率表现相对不足。这是因为模型学习少数类的难度较大，而增量学习场景更是增大了这一难度。模型无法有效识别少数类，极大地影响了平均计算所有类别的精确率和召回率所得的整体结果。同时，受限于模型自身的拟合能力，RNN 模型对少数类的识别能力较差。尽管如此，其性能依然优于基线模型。所提方法在 CNN 模型和 DNN 模型上也同样取得了整体优于基线模型的性能。总体而言，所提方法通过增强模型特征提取能力以及缓解灾难性遗忘与分类偏向问题，有效提升了增量式 IDS 的识别性能。

表 3　CLRD-IIDM 多模型对比实验结果

Tab. 3　Comparative results of CLRD-IIDM on multiple models

模型	任务阶段	指标		
		P/%	*R*/%	*A*/%
基线(CNN)	任务 1	48.51	48.58	75.92
	任务 2	33.72	35.51	61.70
	任务 3	34.74	32.69	58.52
CLRD-IIDM(DNN)	任务 1	50.10	42.29	74.65
	任务 2	39.85	36.50	68.07
	任务 3	40.92	36.17	67.04
CLRD-IIDM(CNN)	任务 1	51.56	43.59	77.62
	任务 2	44.85	37.70	64.63
	任务 3	42.93	42.31	60.63
CLRD-IIDM(RNN)	任务 1	54.39	54.12	73.49
	任务 2	37.90	34.03	64.55
	任务 3	35.31	40.08	62.13
CLRD-IIDM(CNN-GRU)	任务 1	52.52	51.05	81.29
	任务 2	42.91	37.84	68.68
	任务 3	51.99	43.62	68.30

5.5　对比学习增强方法实验

为探究对比学习增强方法在 CLRD-IIDM 的效用，本文以 CNN-GRU 为基准模型开展实验。实验在以下 4 种条件下进行：1）有投影分支并采用旧类补偿对比损失（所提方法），2）有投影分支并采用传统对比损失（无旧类补偿对比损失），3）无投影分支并采用旧类补偿对比损失（无投影分支），4）无对比学习增强方法。需要注意的是，当不采用投影分支时，旧类补偿对比损失直接应用于特征提取层。实验结果如图 3 所示。

由图 3 可知，与传统对比损失相比，旧类补偿对比损失通过调整新旧类样本学习权重使得模型在 3 个任务中准确率分别提升了 3.12%、3.32%、5.61%。在 3 个增量学习任务中，对比学习增强方法中旧类补偿对比损失的应用显著提高了模型检测的准确率。尽管任务 1 不存在旧类样本，不需要调

整模型对新旧类的学习速度,但旧类补偿对比损失仍可以根据模型对任务 1 中各类别样本的拟合情况平衡不同样本的学习权重,表明即使在非增量场景下,对比学习增强方法同样有助于提升模型性能;模型在任务 2、任务 3 中的表现则说明在新旧类样本数量失衡的情况下,旧类补偿对比损失能有效保持模型对旧类的特征提取能力。此外,观察采用所提方法与无投影分支的结果,在任务 3 中,所提方法使模型准确率提升了 9.63%。而在仅采用旧类补偿对比损失、未引入投影分支的条件下,准确率仅提升 3.61%,说明投影分支对数据原始特征的低维空间映射有助于模型学习各类别更加有效的特征表示,投影分支的应用进一步增强了旧类补偿对比损失对模型性能的提升效果。

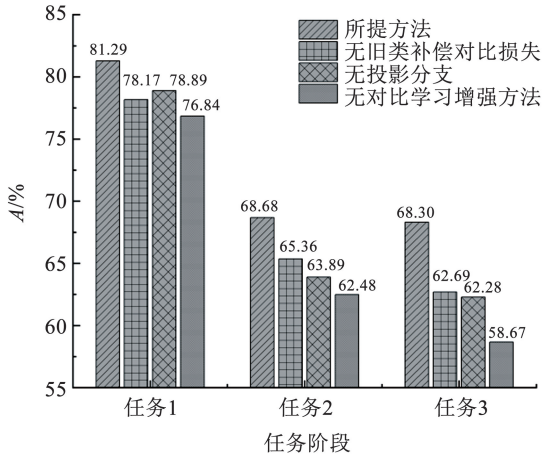


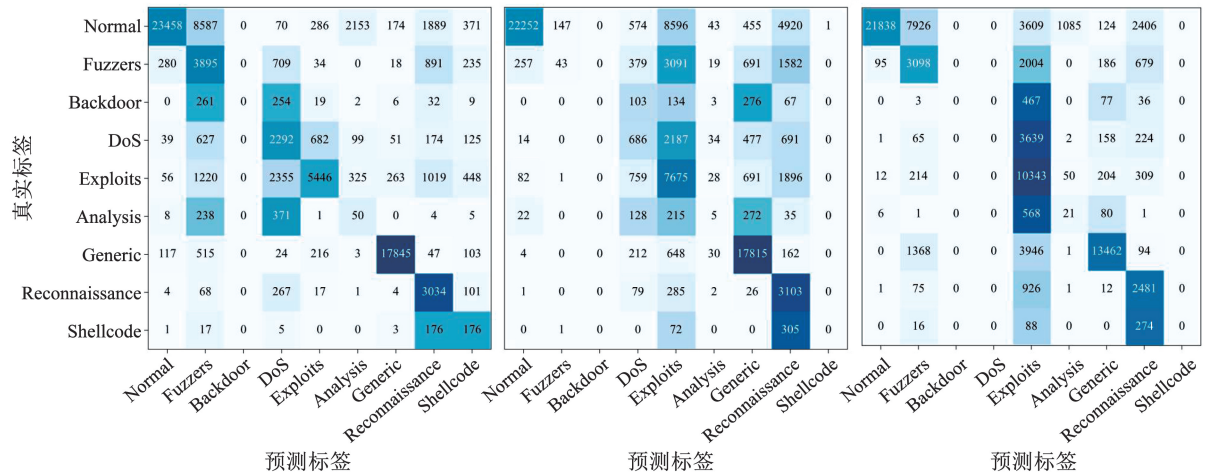
图 3 对比学习增强方法对模型性能的影响

Fig. 3 Impact of contrastive learning enhancement method on model performance

图 4 给出了不同条件下 CNN-GRU 在 UNSW-NB15 测试集上的混淆矩阵,以进一步具体分析投影分支和旧类补偿对比损失对模型识别能力的影响。

对比图 4(a)和图 4(b)可知,采用旧类补偿对比损失的模型在旧类识别上的性能显著优于采用传统对比损失的模型。在采用旧类补偿对比损失的条件下,模型对所有旧类的识别准确率为 59.02%,而改为传统对比损失后,模型对所有旧类的识别准确率下降为 51.50%,且后者在任务 1 中 3 个旧类上的识别准确率相比于前者下降了 11.59%,表明在增量学习过程中,旧类补偿对比损失能有效帮助模型维持对旧类知识的记忆,缓解灾难性遗忘问题,而去除旧类补偿对比损失后,模型会随着增量任务的更新逐渐加重对早期旧类知识的灾难性遗忘。此外,采用旧类补偿对比损失的模型对于任务 3 的 3 个新类的识别准确率为 92.57%,而采用传统对比损失的模型对这 3 个新类的识别准确率为91.97%,表明旧类补偿对比损失对各类样本学习速度的平衡并不会造成对新类识别能力的下降。

对比图 4(a)和图 4(c)可知,投影分支的缺失削弱了旧类补偿对比损失对数据特征的优化能力,导致模型对数据的识别偏向于 Exploits 类,同时对其他各类的识别水平出现了不同程度的下降。具体分析,投影分支通过一系列非线性变换将数据原始特征映射至低维空间,这些变换可以提取更具判别力的特征,同时压缩冗余信息,提升对比损失对数据特征的优化效果。实验结果表明,在对比学习增强方法中引入投影分支是必要的。



(a) 所提方法条件下的结果 (b) 无旧类补偿对比损失条件下的结果 (c) 无投影分支条件下的结果

图 4 不同条件下 CNN-GRU 在 UNSW-NB15 测试集上的混淆矩阵

Fig. 4 Confusion matrix of CNN-GRU on the UNSW-NB15 test set under different conditions

5.6 新旧类关系蒸馏方法实验

为探究新旧类关系蒸馏方法对模型识别能力的具体影响,将完整所提方法(CLRD-IIDM)与消融掉新旧类关系蒸馏方法(CL-IIDM)应用于 CNN-GRU

模型的性能进行对比，结果见表 4。

表 4　新旧类关系蒸馏方法消融实验结果

Tab. 4　Ablation results of the new and old class relational distillation method

方法	任务阶段	指标		
		P/%	*R*/%	*A*/%
CL-IIDM	任务 1	52.52	51.05	81.29
	任务 2	34.72	29.86	65.01
	任务 3	43.52	36.10	64.20
CLRD-IIDM	任务 1	52.52	51.05	81.29
	任务 2	42.91	37.84	68.68
	任务 3	51.99	43.62	68.30

由表 4 可知，新旧类关系蒸馏方法消融与否不会影响任务 1 的结果，因为任务 1 中仅包含新类数据，模型训练过程中不涉及新旧类关系蒸馏方法。而在任务 2、任务 3 中，新旧类关系蒸馏方法的应用使得模型在各项指标上均获得明显的提升，这是因为新旧类关系蒸馏方法利用类关系标签对模型施加正则化约束，减少了模型对新类的分类偏向，提升了模型对旧类的识别。

图 5 和图 6 给出了采用完整所提方法的模型和消融掉新旧类关系蒸馏方法后的模型在 UNSW-NB15 测试集上的混淆矩阵，进一步对比分析新旧类关系蒸馏方法在模型增量训练过程中的具体效用。

由图 6 可以看出，消融掉新旧类关系蒸馏方法后，模型失去了类关系标签对新类学习的正则化约束，对各类别样本的预测大部分偏向 Reconnaissance 类，而对 Exploits 类和 DoS 类的错误预测最为明显，达到各自类别数量的 36.86% 和 70.40%，模型在增量学习任务中出现了明显的分类偏向问题。

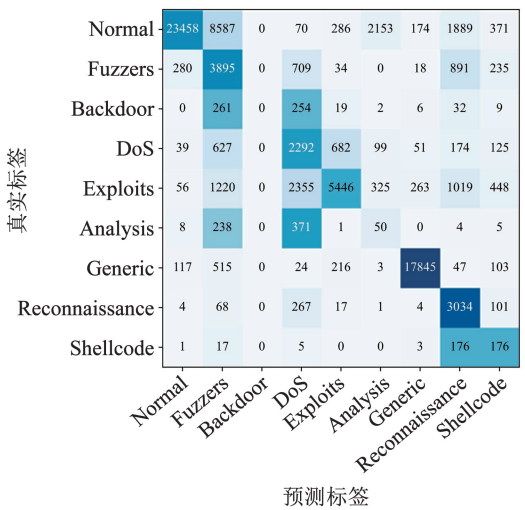


图 5　采用完整所提方法的模型在测试集上的混淆矩阵

Fig. 5　Confusion matrix of the model using the complete proposed method on the test set

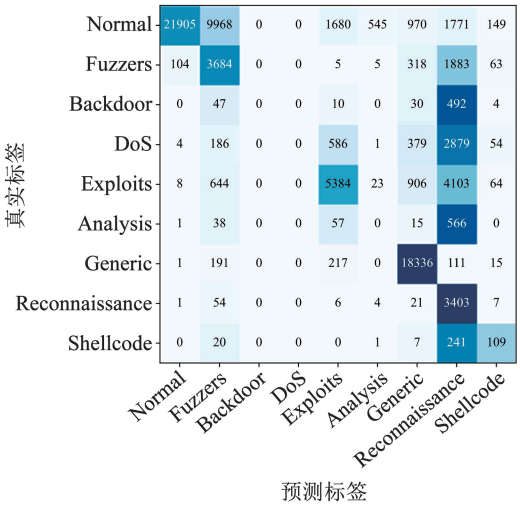


图 6　消融新旧类关系蒸馏方法后模型在测试集上的混淆矩阵

Fig. 6　Confusion matrix of the model on the test set after ablation of the new and old class relational distillation method

对比图 5 和图 6 可以看出，消融掉新旧类关系蒸馏方法后，模型对新类的识别准确率提升了 3.49%，对于两个历史增量任务中的旧类，其识别准确率分别下降了 4.04% 和 15.12%。尽管模型在新类识别上有所改善，但这些改进主要是分类偏向问题的附加效果导致的。新旧类关系蒸馏方法中的类关系标签可以帮助模型在学习新类的同时关注新旧类之间的相似关系，消融该方法后导致模型无法有效保持对旧类知识的记忆，进而导致模型对旧类知识的遗忘。实验结果表明，新旧类关系蒸馏方法有效缓解了模型的分类偏向问题，在保证新类检测准确率无明显变化的情况下，提升了对旧类的识别能力。

5.7　均衡偏差示例存储器实验

为验证均衡偏差示例存储器在 CLRD-IIDM 中的效用，以 CNN-GRU、DNN、RNN 为基准模型，在任务 1 训练结束的基础上，分别采用以下方式更新示例存储器：1）采用均衡偏差示例选取方法（存储均衡样本），2）选取偏差距离最小和最大的样本进行存储（存储非均衡样本），3）随机选取样本进行存储（存储随机样本）。随后在任务 2、任务 3 中对比存储器中不同样本类型对模型性能的影响，结果如图 7所示。

由图 7 可以看出，存储均衡样本使得 CNN-GRU 模型在任务 3 上的准确率比存储非均衡样本提高了 5.40%。在相同的对比条件下，DNN 模型在任务 3 中的准确率提高了 4.95%，RNN 模型在任务 3 中的准确率提高了 3.89%。尽管在任务 2 中存储均衡

样本相比存储非均衡样本对 3 个模型的优势不如在任务 3 中明显,但准确率仍然提升了 2.35% ~3.46%。尽管偏离中心的样本有助于提升对比学习增强方法的效果,但存储偏差距离最小和最大的样本使得存储的样本类型过于单一,且数据集中存在一些位于数据分布边缘的异常样本,其对模型性能的提升是有害的。存储偏差距离最小和最大的样本的方法将这些位于数据分布边缘的异常样本过多地存入存储器中,在模型训练时,影响了对比学习增强方法对数据特征的提取能力,降低了模型的检测性能。与之相反,均衡偏差示例选取方法从整个类别的数据分布中选取旧类样本,避免异常样本过多存储的同时,提高了存储样本的特征表征全面性。此外,对比存储均衡样本与存储随机样本的结果可知,在 3 个模型上,存储均衡样本的准确率均高于存储随机样本,说明均衡偏差示例选取方法选取的样本更有效地表征了旧类特征,模型记忆到了更多的旧类知识。综上,存储均衡样本的均衡偏差示例存储器能保证回放样本对旧类数据有效全面的特征表征,更有助于模型对旧类知识的保持。

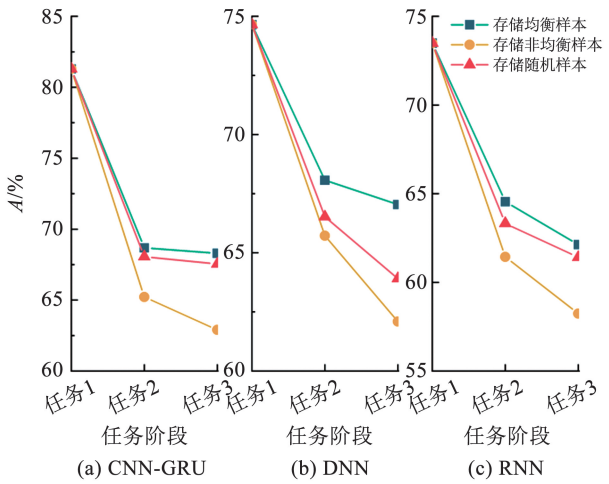


图 7 示例存储器中不同样本类型对模型性能的影响

Fig. 7 Impact of different sample types in the sample memory on the model performance

6 结 论

1)本文提出了基于对比学习和关系蒸馏的增量式入侵检测方法,由对比学习增强方法、新旧类关系蒸馏方法和均衡偏差示例存储器 3 部分构成。

2)对比学习增强方法通过双分支网络结构和旧类补偿对比损失增强了 IDS 的特征提取能力,缓解了 IDS 的灾难性遗忘问题。

3)旧类关系蒸馏方法能够在模型训练时对新类学习施加正则化约束,有效缓解了模型的分类偏向问题。

4)均衡偏差示例存储器通过存储并回放有效全面表征旧类特征的旧类样本,有效保持了模型对旧类攻击的泛化能力。

5)在多个神经网络模型上的对比实验与消融实验显示,应用本文所提方法后,模型在学习新类攻击样本的同时更有效地保持了对旧类的识别能力,表明了本文所提方法的泛化性和有效性,为设计、应用于现代网络环境的 IDS 提供了可靠参考。

参考文献

[1]丁红卫, 万良, 龙廷艳. 深度自编码网络在入侵检测中的应用研究[J]. 哈尔滨工业大学学报, 2019, 51(5): 185
Ding Hongwei, Wan Liang, Long Tingyan. Research on the application of deep auto-encoder network in intrusion detection[J]. Journal of Harbin Institute of Technology, 2019, 51(5): 185. DOI: 10.11918/j.issn.0367-6234.201809115

[2]Shankar S S, Hung B T, Chakrabarti P, et al. A novel optimization based deep learning with artificial intelligence approach to detect intrusion attack in network system[J]. Education and Information Technologies, 2024, 29(4): 3859. DOI:10.1007/s10639-023-11885-4

[3]Yi Tao, Chen Xingshu, Zhu Yi, et al. Review on the application of deep learning in network attack detection[J]. Journal of Network and Computer Applications, 2023, 212: 103580. DOI:10.1016/j.jnca.2022.103580

[4]Kirkpatrick J, Pascanu R, Rabinowitz N, et al. Overcoming catastrophic forgetting in neural networks[J]. Proceedings of the National Academy of Sciences of the United States of America, 2017, 114(13): 3521. DOI:10.1073/pnas.1611835114

[5]Gido M V D V, Tinne T, Tolias A S. Three types of incremental learning[J]. Nature Machine Intelligence, 2022, 4(12): 1185. DOI:10.1038/s42256-022-00568-3

[6]王一丰, 郭渊博, 陈庆礼, 等. 基于对比增量学习的细粒度恶意流量分类方法[J]. 通信学报, 2023, 44(3): 1
Wang Yifeng, Guo Yuanbo, Chen Qingli, et al. Method based on contrastive incremental learning for fine-grained malicious traffic classification[J]. Journal on Communications, 2023, 44(3): 1. DOI:10.11959/j.issn.1000-436x.2023068

[7]Amalapuram S K, Channappayya S S, Tamma B R. Augmented memory replay-based continual learning approaches for network intrusion detection[J]. Advances in Neural Information Processing Systems, 2023, 36: 17156. DOI:10.5555/3666122.3666872

[8]Du Lei, Gu Zhaoquan, Wang Ye, et al. A few-shot class-incremental learning method for network intrusion detection[J]. IEEE Transactions on Network and Service Management, 2024, 21(2): 2389, DOI:10.1109/TNSM.2023.3332284

[9]Liu Huan, Gu Li, Chi Zhixiang, et al. Few-shot class-incremental learning via entropy-regularized data-free replay[C]//17th European

Conference on Computer Vision. Tel Aviv: Springer, 2022: 146. DOI:10.1007/978-3-031-20053-3_9

[10] Shi Yanyan, Shi Dianxi, Qiao Ziteng, et al. Multi-granularity knowledge distillation and prototype consistency regularization for class-incremental learning[J]. Neural Networks, 2023, 164: 617. DOI:10.1016/j.neunet.2023.05.006

[11] Korycki L, Krawczyk B. Class-incremental experience replay for continual learning under concept drift [C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. Nashville: IEEE, 2021: 3649. DOI: 10.1109/CVPRW53098.2021.00404

[12] Li Yichen, Li Qunwei, Wang Haozhao, et al. Towards efficient replay in federated incremental learning [C]//2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2024: 12820. DOI:10.1109/CVPR52733.2024.01218

[13] Lin Huiwei, Feng Shanshan, Li Xutao, et al. Anchor assisted experience replay for online class-incremental learning[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2023, 33(5): 2217. DOI:10.1109/TCSVT.2022.3219605

[14] Zhu Fei, Cheng Zhen, Zhang Xuyao, et al. Class-incremental learning via dual augmentation[J]. Advances in Neural Information Processing Systems, 2021, 17: 14306. DOI:10.5555/3540261.3541357

[15] Chen Ting, Kornblith S, Norouzi M, et al. A simple framework for contrastive learning of visual representations [C]//37th International Conference on Machine Learning. Vienna: IMLS, 2020: 1597. DOI:10.5555/3524938.3525087

[16] 刘雨薇, 陈松灿. 基于 IMIU 的在线类增量对比学习[J]. 软件学报, 2024, 35(12): 5544
Liu Yuwei, Chen Songcan. Online class incremental contrastive learning based on incremental mixup-induced universum [J]. Journal of Software, 2024, 35(12): 5544. DOI: 10.13328/j.cnki.jos.007094

[17] Khosla P, Teterwak P, Wang Chen, et al. Supervised contrastive learning[J]. Advances in Neural Information Processing Systems, 2020, 33: 18661. DOI:10.5555/3495724.3497291

[18] Smith J, Hsu Y C, Balloch J, et al. Always be dreaming: a new approach for data-free class-incremental learning[C]//18th IEEE/CVF International Conference on Computer Vision. Montreal: IEEE, 2021: 9354. DOI:10.1109/ICCV48922.2021.00924

[19] Bang J, Kim H, Yoo Y J, et al. Rainbow memory: continual learning with a memory of diverse samples[C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021: 8218. DOI:10.1109/CVPR46437.2021.00812

[20] Moustafa N, Slay J. UNSW-NB15: a comprehensive data set for network intrusion detection systems(UNSW-NB15 network data set) [C]//2015 Military Communications and Information Systems Conference. Canberra: IEEE, 2015: 7348942. DOI: 10.1109/MilCIS.2015.7348942

[21] Vinayakumar R, Alazab M, Soman K P, et al. Deep learning approach for intelligent intrusion detection system [J]. IEEE Access, 2019, 7: 41525. DOI:10.1109/ACCESS.2019.2895334

[22] Vinayakumar R, Soman K P, Prabaharan P. Applying convolutional neural network for network intrusion detection[C]//2017 International Conference on Advances in Computing, Communications and Informatics. Manipal: IEEE, 2017: 1222. DOI:10.1109/ICACCI.2017.8126009

[23] Yin Chuanlong, Zhu Yuefei, Fei Jinlong, et al. A deep learning approach for intrusion detection using recurrent neural networks [J]. IEEE Access, 2017, 5: 21954. DOI:10.1109/ACCESS.2017.2762418

[24] Cao Bo, Li Chenghai, Song Yafei, et al. Network intrusion detection model based on CNN and GRU[J]. Applied Sciences-Basel, 2022, 12(9): 4184. DOI:10.3390/app12094184