2025 年 第 25 卷 第 19 期
2025, 25(19):08142-09

科 学 技 术 与 工 程
Science Technology and Engineering

ISSN 1671—1815
CN 11—4688/T

DOI:10.12404/j.issn.1671-1815.2405694

引用格式:闫计栋, 钟美荟, 周帆. TopoSMOTE: 基于拓扑数据分析的网络入侵检测不平衡学习[J]. 科学技术与工程, 2025, 25(19): 8142-8150.

Yan Jidong, Zhong Meihui, Zhou Fan. TopoSMOTE: topological data analysis-based imbalanced learning for network intrusion detection[J]. Science Technology and Engineering, 2025, 25(19): 8142-8150.

TopoSMOTE: 基于拓扑数据分析的网络入侵检测不平衡学习

闫计栋[1], 钟美荟[2], 周帆[2*]

(1. 国家能源集团科技与信息化部, 北京 100011; 2. 电子科技大学信息与软件工程学院, 成都 610054)

摘 要 网络入侵检测系统(network intrusion detection systems,NIDS)对维护网络安全至关重要。然而,由于网络流量数据的复杂性和类不平衡问题,现有检测模型往往出现高误报率和对不同攻击类型的检测精度不足的现象。为了克服这些挑战,提出了一种基于拓扑数据分析(topological data analysis,TDA)的网络入侵检测不平衡学习方法,称为 TopoSMOTE,用于生成新的少数类以平衡训练样本。TopoSMOTE 的核心在于构建拓扑图来合成新样本。首先,该方法使用 TDA 映射网络流量数据中的空间关系和连接模式,并构建拓扑图。然后,基于拓扑图设计了一种少数类样本选择策略,通过低维映射空间中的距离度量选择具有拓扑关系的最近邻样本来合成新数据。本文在两个类不平衡的数据集上进行了实验。实验结果表明,与先进的过采样方法和入侵检测模型相比,TopoSMOTE 方法具有更高的检测精度和更低的误报率。

关键词 网络入侵检测; 拓扑数据分析; 不平衡学习; 数据增强

中图法分类号 TP309; **文献标志码** A

TopoSMOTE: Topological Data Analysis-based Imbalanced Learning for Network Intrusion Detection

YAN Ji-dong[1], ZHONG Mei-hui[2], ZHOU Fan[2*]

(1. Technology and Information Department of National Energy Group, Beijing 100011, China;
2. School of Information and Software Engineering, University of Electronic Technology of China, Chengdu 610054, China)

[**Abstract**] Network intrusion detection systems (NIDS) are critical for maintaining cybersecurity. However, due to the complexity of network traffic data and the issue of class imbalance, existing detection models often exhibit high false alarm rates and insufficient detection accuracy for different types of attacks. To address these challenges, an imbalanced learning method for network intrusion detection, based on topological data analysis (TDA) and named TopoSMOTE, was proposed. This method aims to balance the training dataset by generating new minority class samples. The core of TopoSMOTE lied in constructing topological graphs to synthesize new samples. Firstly, the method used TDA to map the spatial relationships and connection patterns in network traffic data, forming a topological graph. Then, based on the topological graph, a minority class sample selection strategy was designed, which synthesized new data by selecting the nearest neighbor samples with topological relationships in a low-dimensional mapped space. Experiments were conducted on two imbalanced datasets. The experimental results show that the TopoSMOTE method achieves higher detection accuracy and lower false alarm rates compared to advanced oversampling methods and intrusion detection models.

[**Keywords**] network intrusion detection; topological data analysis; imbalanced learning; data augmentation

随着网络流量的持续增长和网络攻击类型的日益复杂化,网络入侵检测系统(network intrusion detection system, NIDS)在维护网络安全方面变得愈发重要[1]。这些系统能够识别出各种恶意活动,如病毒传播、黑客入侵和其他形式的未授权访问,并能够及时响应以防御这些活动,在很大程度上保护

收稿日期: 2024-07-29 **修订日期**: 2024-12-23

基金项目: 国家自然科学基金(62072077, 62176043);四川省科技计划(2021YFQ0007);四川省自然科学基金(2022NSFSC0505)

第一作者: 闫计栋(1983—),男,汉族,山西吕梁人,博士研究生,高级工程师。研究方向:人工智能,电力信息化,网络安全。E-mail:16810080@ceic.com。

* **通信作者**: 周帆(1981—),男,汉族,四川眉山人,博士,教授。研究方向:机器学习,时空数据挖掘,数据挖掘与知识发现。E-mail:fan.zhou@uestc.edu.cn。

了网络安全[2]。

为了提高 NIDS 的准确性和适应性,研究人员引入了不同的机器学习和深度学习技术[3-4]。这些技术能够处理大规模的网络流量数据,提取有意义的行为模式,从而提升模型的检测能力。尽管通过不断改进和优化网络结构可以提升模型性能,但这种提升的幅度正在缩小。一个主要挑战在于,面对大规模的高维流量数据,基于卷积神经网络的入侵检测模型往往难以捕获数据类别之间的相互关系和拓扑结构[5],如流量的时间序列特征和网络的拓扑结构,这限制了它们在处理复杂网络环境时的性能。另一个主要挑战是类不平衡问题,即正常行为的数据点远多于异常行为的数据点。这种不平衡导致检测模型倾向于学习多数类的特征分布,而忽略少数的潜在攻击行为,从而导致较高的假阴性率[6]。为了应对类不平衡问题,研究者提出了多种策略。常见的方法包括过采样[7-8]和欠采样[9]。例如,合成少数类过采样技术(synthetic minority oversampling technique, SMOTE)通过在特征空间中创建新的少数类实例,增加少数类的样本量,从而改善模型的学习过程。SMOTE 的关键优点在于它通过在少数类样本之间插值生成新样本,保持数据的多样性,避免了简单地复制少数类样本而导致过拟合问题。然而,直接在原始输入数据中进行插值过采样可能会生成域外样本,这可能会影响模型的准确性[10]。欠采样方法通过减少多数类样本的数量来平衡类分布,但可能会丢失重要信息。

此外,随机森林和 Boosting 等集成方法也被用于提高对少数类的识别能力。文献[11]提出融合随机森林模型来进行特征选择,并且使用梯度提升决策树模型来实现分类,有效地缓解了类不平衡的问题。文献[12]构建了一个动态集成模型,将动态欠采样机制融入 Boosting 框架中以应对欠采样导致的噪声样本过拟合问题。深度学习方法也被应用于解决类不平衡问题。例如,使用成本敏感的损失函数[13-14]来优化模型对少数类的学习过程。这些方法在处理类不平衡问题方面表现出有效性,但在网络安全领域的应用仍面临挑战,特别是在保持数据内在结构和关系完整性方面这些方法还略显不足[15]。

为了应对这些挑战,提出 TopoSMOTE 方法,这是一种基于拓扑数据分析(topological data analysis, TDA)的入侵检测类别不平衡学习方法。TopoSMOTE 将数据的拓扑结构纳入分析维度,有助于更深入地分析网络流量的分布模式[16-17]。为了让增强后的样本保留原有的拓扑连接信息,设计针对拓扑图的少数类选取策略。该方法考虑到拓扑结构的特性,将类不平衡学习融入到拓扑结构,进而有效处理网络环境中正常行为与异常行为之间由于数量不均衡导致的类不平衡问题。

TopoSMOTE 方法在解决网络流量类别不平衡问题方面具有重要的理论价值和实践意义。对于企业应用,该方法可以有效提升关键基础设施领域的网络安全防护能力,优化网络流量监控技术,为网络安全产品的研发提供创新技术支撑。在网络攻击发生时,TopoSMOTE 可以指导工程设计更精准的响应机制,从而提升应急响应效率。

1　相关工作

针对入侵检测数据不平衡问题,解决方法主要分为两个方面:基于数据层面的方法和基于算法层面的方法[18]。

基于数据层面的方法主要通过数据抽样技术调整各类别比例以构建模型。文献[19]通过结合深度学习方法和统计思想来解决少数样本攻击检测的问题,具体提出一种基于改进条件变分自编码器(improved conditional variational autoencoder, improved-CVAE)和边界合成少数过采样技术的物联网入侵检测方法。该方法通过在 CVAE 中引入辅助网络来调整编码器的输出概率分布,学习不同类别样本的后验分布,使同类样本的分布集中,不同类别样本的分布在嵌入空间中分散。然后在嵌入空间中对少数样本进行边界过采样,自适应地生成代表性的样本,以平衡数据集。文献[20]提出了一种改进的混合采样模型,使用自适应合成采样(adaptive synthetic sampling, ADASYN)生成少数类样本,并使用高斯混合模型(Gaussian mixture model, GMM)进一步生成新样本。

基于算法层面的方法主要通过改进算法的训练过程及采用多种集成策略来提升模型性能。文献[21]提出了一种基于 Siamese 神经网络的入侵检测系统(Siamese intrusion detection system, Siam-IDS),用于解决入侵检测系统中的类别不平衡问题。Siam-IDS 通过计算输入对之间的相似度来区分相似和不相似的样本,而非使用传统的过采样和随机欠采样方法。文献[22]提出了一种基于重采样随机森林(resample-random forest, Resample-RF)的物联网入侵检测方法,旨在使用集成策略来解决不平衡样本问题。文献[23]提出了一种基于 K 近邻(K-nearest neighbor, KNN)和生成对抗网络(generative adversarial networks, GAN)的混合方法。方法主要使用了一种改进的表格辅助分类生成对抗网络(table auxiliary classifier GAN, TACGAN)生成更

逼真的攻击数据,最终将处理后的正常数据和攻击数据混合,形成平衡的数据集。

近年来,TDA[24]作为一种结合代数拓扑和计算几何的创新方法,在降低数据维度和挖掘数据集中的潜在模式方面表现出色。TDA 为研究者提供了独特的视角,其有助于揭示并分析复杂数据的相互关系和整体结构[25-26]。TDA 的应用范围广泛,涵盖图像和信号处理、机器人技术等[27-29]。在对数据变化的鲁棒性方面,TDA 也表现出显著优势[16,30],即使在存在大量数据噪声或不完整的情况下,TDA 仍能提供可靠信息。这在处理如网络流量数据等受多种因素影响的数据时尤为重要,因为这类数据常受到如网络延迟、数据包丢失等影响。文献[31]使用 TDA 和无监督学习来检测针对基于机器学习的网络入侵检测系统的数据投毒攻击。

2 TopoSMOTE 方法

本研究探讨了一种新颖的方法,即在流量数据构建的拓扑结构中合成少数类样本以平衡训练数据。尽管从网络流量数据中生成的拓扑图在一定程度上体现了数据的内在联系和模式,但这些图中的节点表示仍然存在类别不平衡的问题,某些类型样本在节点中的数量明显少于其他类型。这种不平衡可能导致对攻击样本的学习不够充分,影响检测模型的性能。

大多数先前工作在原始空间中为每个少数类样本合成新样本,本研究的 TopoSMOTE 方法则仅选择具有关联拓扑关系的少数类样本,并使用低维映射空间中的距离作为选择最近邻样本的度量。图 1 所示为 TopoSMOTE 方法的示意图,展示了一个简单的拓扑图,其中表示的蓝色虚线圈表示拓扑图中的节点,节点之间根据特定规则添加了边。

具体来说,TopoSMOTE 在网络流量数据中生成拓扑图的步骤如图 2 所示。首先,算法将原始的高维数据 $\boldsymbol{X}$ 通过过滤函数 f 映射到低维的特征空间 $\boldsymbol{Z}$ 中,表示为 $f:\boldsymbol{X}\rightarrow\boldsymbol{Z}$。通常空间 $\boldsymbol{Z}$ 是一维或者低维的欧式空间。均匀流形近似和投影(uniform manifold approximation and projection, UMAP)[32]是一种有效的维数缩减技术,专门用于捕捉高维数据中的流形结构。文献[33]研究证明了网络流量数据会以黎曼流形的形式出现。这表明 UMAP 凭借其处理流形数据的能力可以成为分析和可视化此类数据的有效工具。因此在本研究中,选择 UMAP 算法作为过滤函数,以便更好地理解数据。

过滤函数考虑了每个点到 x 轴的距离,此处定义交叉熵 C 作为过滤函数的优化目标,表达式为

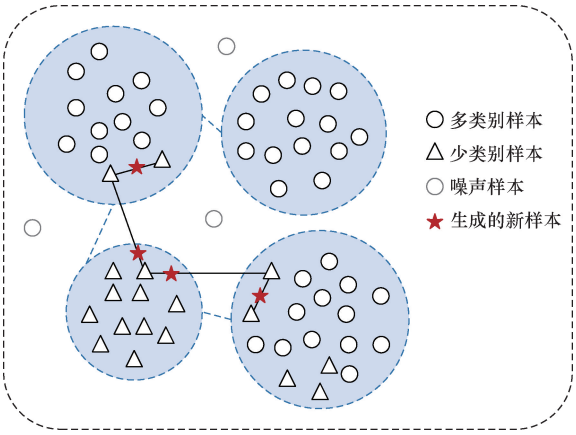

图 1 所提 TopoSMOTE 方法的示意图

Fig. 1 Illustration of the proposed TopoSMOTE

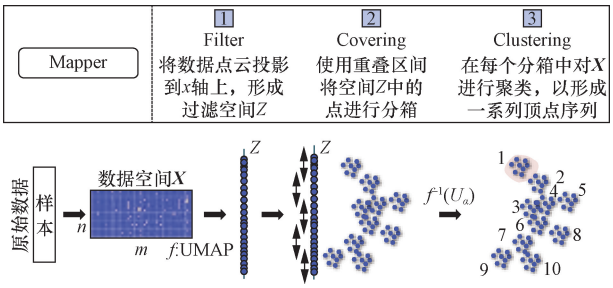

图 2 拓扑图生成过程示意图

Fig. 2 Illustration of topological graph generation process

$$C[w_{\mathrm{h}}(e),w_{\mathrm{l}}(e)]\triangleq\sum_{e\in E}w_{\mathrm{h}}(e)\lg\left[\frac{w_{\mathrm{h}}(e)}{w_{\mathrm{l}}(e)}\right]+[1-w_{\mathrm{h}}(e)]\lg\left[\frac{1-w_{\mathrm{h}}(e)}{1-w_{\mathrm{l}}(e)}\right]\tag{1}$$

式(1)中:通过降维映射后,将原始数据构建为一个拓扑图结构,其中节点表示数据样本,边 $e\in E$ 表示样本之间的连通关系,E 表示边的集合;$w_{\mathrm{h}}(e)$ 和 $w_{\mathrm{l}}(e)$ 分别为高维和低维空间中的权重。具体表示为

$$w_{\mathrm{h}}(e)=\exp\left[-\frac{d(x_i,x_j)-\rho_i}{\sigma_i}\right]\tag{2}$$

式(2)中:ρ_i 为点 x_i 的距离偏移量(通常是指该点到指定数量 k 范围中邻近点的最小距离);σ_i 为用于缩放距离的参数;$d(x_i,x_j)$ 为点 x_i 到点 x_j 的距离。σ_i 和 ρ_i 的关系表示为

$$\sum_{j=1}^{k}\exp\left\{\frac{-\max[0,d(x_i,x_j)-\rho_i]}{\sigma_i}\right\}=\log_2 k\tag{3}$$

$$w_{\mathrm{l}}(e)=[1+a(z_i-z_j)^{2b}]^{-1}\tag{4}$$

式中:z_i 和 z_j 分别为低维空间中的点 i 和 j 的坐标;a 和 b 为调节参数,用于控制两点在低维空间中的距离。这些权重用于优化目标函数 C,以便在高维和低维空间之间保持相似的结构。

接着,在空间 $\boldsymbol{Z}$ 中定义一组有限的覆盖 U =

$\{U_\alpha\}_{\alpha\in A}$,其中每个覆盖 U_α 都是空间 $\boldsymbol{Z}$ 中的子区间,A 为有限索引集。每个覆盖是等长且具有间隔的,每个间隔之间存在 $p\%$ 的重叠。具体的重叠百分比 p 计算公式为

$$p = \frac{rl - n}{rl - l} \tag{5}$$

式(5)中:r 为覆盖的固定长度;l 为划分的间隔数;n 为样本总数。

由于函数 f 是连续的,因此集合 $f^{-1}(U_\alpha)$ 形成了 $\boldsymbol{X}$ 的一个开覆盖。在 $\boldsymbol{X}$ 的每个覆盖 $f^{-1}(U_\alpha)$ 内部,数据点基于原始数据的相似性进行聚类。这一步骤中,每个聚类可视为 $\boldsymbol{X}$ 中的一组子集。对于每个子集 α 内的数据点集 $f^{-1}(U_\alpha)$ 可以进一步细分为 $\bigcup_{i=1}^{k_\alpha} V(\alpha,i)$,记作 $f^{-1}(U_\alpha) = \bigcup_{i=1}^{k_\alpha} V(\alpha,i)$,其中 V 为聚类中心,k_α 为 $f^{-1}(U_\alpha)$ 中聚类数量。

最后,基于网络流量数据的拓扑图和原始数据特征构建拓扑图 $\boldsymbol{W}$。拓扑图 $\boldsymbol{W}$ 中的每个元素 $\boldsymbol{W}(x_v, x_u)$ 描述了样本点之间的拓扑关系。表达式为

$$\boldsymbol{W}(x_v,x_u) = \boldsymbol{1}_{V(x_v)=V(x_u)} + \boldsymbol{1}_{V(x_v)\neq V(x_u)\wedge V(x_v)\sim V(x_u)} \tag{6}$$

式(6)中:$V(x_v)$ 和 $V(x_v)$ 表示点 x_u 和 x_v 所在的聚类,$\boldsymbol{1}$ 为指示函数。公式表示当两点处在同一聚类中或者在不同聚类中但相连的时取值为 1,否则为 0。

对于少数类样本的选择,TopoSMOTE 在生成的拓扑图中从少数类节点中随机选择一部分节点作为基础节点,特别是选择那些靠近类别边界的节点。因为这些节点的信息量更大,生成的样本更有助于分类器的学习[34]。其中类别边界的定义是如果节点中多数类样本数量超过节点大小一半,则认为是类别边界节点。对于每个被选中的基础节点中的样本,找到其在节点中的其他最近邻样本(同类样本),查找最近邻样本公式为

$$\mathrm{nn}(v) = \arg\min_{u\in S} \| h_v - h_u \| \tag{7}$$

式(7)中:$\mathrm{nn}(v)$ 表示样本点 x_v 的最近邻样本;S 为同类样本集合;h_v 和 h_u 分别表示样本点 x_v 和 x_u 在低维空间中的表示。

为了更好地利用拓扑图,不仅在每个节点中进行过采样,还在两个有连接的节点中选取少数样本进行采样。这些节点之间的连接反映了数据点之间的重要关系,通过在这些节点之间进行采样,可以更有效地保留拓扑图中的关键关系。

随后使用 SMOTE 算法,对基础样本及其最近样本进行插值,生成新的少数类样本,最终生成增强的拓扑图,插值公式为

$$x'_v = (1-\delta)x_v + \delta x_{nn(v)} \tag{8}$$

式(8)中:x'_v 为新生成的样本表征;$x_{\mathrm{nn}(v)}$ 为最近邻节点的表征;δ 为 0~1 的随机数。

所提方法的核心在于,在已有的拓扑结构上生成新的少数类样本,同时保持图的结构特征和节点间的拓扑关系。在这个中间映射空间中,维度比原始维度低,因此来自同一类别的样本分布会更加密集。由于类内相似性以及类间差异已经被拓扑结构固定,插值可以更可靠地生成域内样本。

3　实验分析

3.1　数据集特性与数据预处理

本文中使用两个不平衡的公开数据集 CIC-IDS2017[35] 和 CIC-MalMen2022[36]。对于数据划分,随机选择 75% 的流量用于训练,剩下的 25% 用于测试。

(1)CIC-IDS2017 数据集。CIC-IDS2017 数据集包含与真实世界数据相似的良性流量和最新的常见攻击类型。数据集共包含 3 119 345 个实例和 83 个特征,其中包括 1 个类别标签和 82 个流量特征,标签包括良性流量和 14 种现代网络攻击。

(2)CIC-MalMen2022 数据集。该数据集由间谍软件、勒索软件和木马恶意软件组成,提供了一个与现实世界中的恶意软件行为相近的数据集。恶意记录进一步细分为 15 种不同的攻击类型。每条记录包括 55 个特征,这些特征中包含了 26 个新的基于内存的特征,这些特征是通过特征提取器生成的。

表 1 和表 2 分别展示了两个数据集的样本分布情况。显然,对于 CIC-IDS2017 数据集,正常行为样本占比超过 80%。相比之下,CIC-MalMen2022 数据集提供了更为平缓的样本分布。

表 1　CIC-IDS2017 数据集的样本分布情况

Table 1　Sample distribution of CIC-IDS2017 dataset

类别	数量
BENIGN	2 273 097
DoS Hulk	231 073
PortScan	158 930
DDoS	128 027
DoS GoldenEye	10 293
FTP-Patator	7 938
SSH-Patator	5 897
DoS slowloris	5 796
DoS Slowhttptest	5 499
Bot	1 966
Brute Force	1 507
XSS	652
Infiltration	36
Sql Injection	21
Heartbleed	11

表 2 CIC-MalMen2022 数据集的样本分布情况

Table 2 Sample distribution of CIC-MalMen2022 dataset

类别	数量
Normal	29 298
SpywareTransponder	2 410
SpywareGator	2 200
RansomwareShade	2 128
RansomwareAko	2 000
Spyware180solutions	2 000
SpywareCWS	2 000
TrojanRefroso	2 000
TrojanScar	2 000
RansomwareConti	1 988
TrojanEmotet	1 967
RansomwareMaze	1 958
TrojanZeus	1 950
RansomwarePysa	1 717
TrojanReconyc	1 570
SpywareTIBS	1 410

在模型训练之前,所有数据集都经过了预处理。包括去除重复的标题和常数列,空白数据用“0”填充,以及使用 One-Hot 编码将字符特征转换为数值特征。在网络流量数据集中,特征的数值大小差异很大。为了确保特征之间的可比性,每一列都使用了 Z-Score 进行标准化处理。

3.2 评价指标

为了全面评估模型在多类别环境下的表现,一般可采用宏观平均和微观平均两种指标。宏观平均平等对待每一个类别,因此受到攻击类别检测结果的影响较大。微观平均将每个样本视为一个整体,可能会过分强调对正常类别的检测性能。鉴于网络入侵检测中的类别不平衡问题,识别较少见的攻击类型同样重要。因此,本文中选择宏观平均指标来对模型的性能进行综合评估。所使用指标为

$$A = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}} \times 100\% \tag{9}$$

$$P_{\mathrm{macro}} = \frac{1}{N}\sum_{i=1}^{N}\frac{\mathrm{TP}_i}{\mathrm{TP}_i + \mathrm{FP}_i} \times 100\% \tag{10}$$

$$R_{\mathrm{macro}} = \frac{1}{N}\sum_{i=1}^{N}\frac{\mathrm{TP}_i}{\mathrm{TP}_i + \mathrm{FN}_i} \times 100\% \tag{11}$$

$$\mathrm{FPR}_{\mathrm{macro}} = \frac{1}{N}\sum_{i=1}^{N}\frac{\mathrm{FP}_i}{\mathrm{FP}_i + \mathrm{TN}_i} \times 100\% \tag{12}$$

$$F_1\text{-score} = \frac{2P_{\mathrm{macro}}R_{\mathrm{macro}}}{P_{\mathrm{macro}} + R_{\mathrm{macro}}} \times 100\% \tag{13}$$

式中:TP 为真正例,表示模型正确预测为某一类的样本数;TN 为真负例,表示模型正确预测为非某一类的样本数;FP 为假正例,表示模型错误地将其他类预测为该类的样本数;FN 为假正例,表示模型错误地将其他类预测为非该类的样本数;A 为在预测过程中,预测正确的样本数占总样本数的比例。P_{macro}为对每个类别分别计算精确率,然后对所有类别取平均值,称为“宏平均精确率”;R_{macro}为对每个类别分别计算召回率,再对所有类别取平均值,称为“宏平均召回率”;$\mathrm{FPR}_{\mathrm{macro}}$为对每个类别计算假阳性率,即错误地将其他类别预测为该类的比例,然后取平均;F_1-score 为综合考虑精确率和召回率的调和平均值。

3.3 实验设置

为了验证本文所提模型的有效性,实验将 TopoSMOTE 与一些具有代表性的检测模型和不平衡学习方法进行了对比,对比模型如下。

(1)DNN:DNN(deep neural network)是最基本的分类模型,设置 DNN 的目的是为了观察原始数据的检测效果。

(2)AB-LightGBM[37]:一种基于 ADASYN 和贝叶斯优化的 LightGBM 的入侵检测方法。该方法利用 ADASYN 过采样技术来增加训练数据中的攻击样本数量,以解决数据不平衡问题。同时,使用贝叶斯优化对 LightGBM 模型的超参数进行调优,提高检测精度并减少计算负担。

(3)SMOTE-TomekLink[38]:一种结合了 SMOTE-TomekLink 数据平衡技术与机器学习算法的无线传感器网络(wireless sensor networks,WSN)入侵检测方法。该方法通过 SMOTE-TomekLink 技术生成平衡的数据集,并使用随机森林(random forest,RF)算法来提升检测的准确性。

(4)TMG-IDS[39]:一种基于生成对抗网络(generative adversarial network ,GAN)的数据增强模型,用于解决网络入侵检测中的数据不平衡问题。具体而言,所提的 TMG-IDS(transaction monitoring gateway, intrusion detection system)方法通过多生成器结构同时生成不同类型的攻击数据,并引入分类器结构以优化生成器和判别器。生成器损失中加入了生成样本和原始样本,以及生成样本与其他类型生成样本的余弦相似度,从而提高生成样本质量并减少类间重叠。

本文提出的 TopoSMOTE 模型使用 UMAP 作为过滤函数,其参数 a 和 b 分别设置为 1.73 和 0.79。这两个值是文献[32]通过优化过程得出的,旨在找到高斯核与余弦核之间的最佳平衡,以最大程度地在低维嵌入中保留高维空间的拓扑结构。覆盖区间的重叠百分比设置为 40%,r = 30。合成样本的插值参数为 0 ~ 1 的随机数。DNN 模型的详细框架如表 3 所示。

表 3　不同数据集上的 DNN 模型参数

Table 3　DNN model parameters on different datasets

CIC-IDS2017	CIC-MalMen2022
输入数据(78)	输入数据(55)
Reshape(13×6×1)	全连接层(256)
Cov 层 1(13×6×32)	全连接层(128)
MaxPooling 层(1×1)	Dropout 层(0.2)
Flatten 层	全连接层(64)
Dropout 层(0.2)	输出层(15)
全连接层(256)	—
输出层(14)	—

3.4　实验结果与分析

为了验证 TopoSMOTE 方法的有效性,本文进行了大量的实验测试。结果表明,与其他入侵检测方法相比,TopoSMOTE 在检测精度和效率上均有提高,特别是在应对类不平衡问题方面表现出色。

表 4TopoSMOTE 模型与其他模型在 CIC-IDS2017 数据集上对多个攻击类型检测的 F_1-score 的比较结果。从表 4 中可以看出,TopoSMOTE 在大多数攻击类型的检测中均表现优异。例如,在检测 Heartbleed 时, TopoSMOTE 的 F_1-score 达到了 74.51%, 相比于 SMOTE-TomekLink 提升了 63.98%,相比于 TMG-IDS 提升了 20.97%,相比于 AB-LightGBM 提升了 17.97%。这种显著提升主要得益于 TopoSMOTE 能够通过学习少数类攻击样本的分布,生成更多具有代表性的攻击样本,从而提高模型对少数类攻击的检测能力。

由上可知,在 CIC-MalMen2022 数据集上的实验结果更为显著。从表 5 列出的详细对比数据可以看出,TopoSMOTE在检测 SpywareTIBS 时,F_1-score 达到了 97.96%,相比于 TMG-IDS 提升了 30.28%,相比

表 4　不同模型在 CIC-IDS2017 数据集上对多个攻击类型检测的 F_1-score 值比较

Table 4　Comparison of F_1-scores for multiple types detection in CIC-IDS2017 dataset using different models

攻击类型	DNN	AB-LightGBM[25]	SMOTE-TomekLink[26]	TMG-IDS[16]	TopoSMOTE
Bot	0	86.40	88.87	**91.22**	91.02
DDoS	90.84	83.62	87.62	89.90	**91.35**
DoS GoldenEye	90.31	83.35	87.41	90.73	**98.16**
DoS Hulk	49.68	78.65	64.52	80.90	**88.21**
DoS Slowhttptest	0	79.98	91.83	96.3	**98.21**
DoS slowloris	96.71	80.18	76.12	97.45	**98.19**
FTP-Patator	92.54	80.20	88.69	97.00	**98.46**
Heartbleed	0	53.54	10.53	56.13	**74.51**
Infiltration	0	79.96	63.39	95.90	**98.02**
PortScan	96.86	80.06	36.54	97.18	**98.25**
SSH-Patator	96.93	80.03	40.05	97.36	**98.03**
Brute Force	98.25	80.25	60.12	97.35	**98.68**
Sql Injection	0	81.82	71.21	66.67	**82.93**
XSS	65.45	60.91	17.39	72.73	**88.89**

注:最好的结果用粗体显示,次好的结果用下划线表示。

表 5　不同模型在 CIC-MalMen2022 数据集上对多个攻击类型检测的 F_1-score 值比较

Table 5　Comparison of F_1-scores for multiple attack types detection in CIC-MalMen2022 dataset using different models

攻击类型	DNN	AB-LightGBM	SMOTE-TomekLink	TMG-IDS	TopoSMOTE
RansomwareAko	28.67	62.78	41.81	70.28	**76.92**
RansomwareConti	29.54	77.27	60.35	71.21	**86.42**
RansomwareMaze	40.88	81.58	72.45	71.14	**92.69**
RansomwarePysa	28.39	80.77	69.8	69.92	**98.06**
RansomwareShade	39.17	52.33	39.61	50.03	**75.29**
Spyware180solutions	19.62	73.33	44.34	70.58	**78.23**
SpywareCWS	0	74.00	47.76	70.58	**79.79**
SpywareGator	46.87	51.94	37.55	50.59	**74.52**
SpywareTIBS	55.33	73.76	70.48	67.68	**97.96**
SpywareTransponder	33.68	73.18	36.42	53.01	**73.80**
TrojanEmotet	45.21	79.19	66.17	71.24	**89.52**
TrojanReconyc	0	77.52	62.5	68.99	**98.01**
TrojanRefroso	57.82	74.83	51.17	70.58	**81.63**
TrojanScar	45.15	75.88	55.37	70.89	**83.82**
TrojanZeus	35.56	83.60	79.78	70.72	**96.53**

注:最好的结果用粗体显示,次好的结果用下划线表示。

于 SMOTE-TomekLink 提升了 27.48%,相比于 AB-Light GBM 提升了 24.20%。

这种性能的提升不仅体现在某些特定攻击类型上,TopoSMOTE 在整体的检测准确性上也表现出色。表 6 和表 7 分别展示了不同方法在 CIC-IDS2017 和 CIC-MalMen2022 数据集上的整体性能比较。

从表 6 中可以看出,TopoSMOTE 在各个评价指标上都领先于其他方法,除了在准确率指标上略低于 AB-LightGBM 方法 0.16%。原因为TopoSMOTE 生成的特征在少数类检测上的权重较大,导致整体准确率有所降低。

对于更加复杂的 CIC-MalMen2022 数据集来说,各个模型的性能普遍降低,但是所提模型依旧领先于其他模型。从表 6 中可以看出,TopoSMOTE 的准确率达到了 93.41%,精确率为 85.44%,召回率为 86.37%,假阳性率仅为 1.03%。

从以上对比实验结果可以看出,TopoSMOTE 在各个评价指标上均优于其他方法,尤其在应对类不平衡问题上表现出色。这些结果验证了 TopoSMOTE 方法在网络入侵检测中的有效性和鲁棒性。

本节进行了额外的消融实验来验证本文的动机,即在少数类样本选择策略中考虑具有连接边的节点对性能的影响,以及验证使用低维空间中的距离作为选择最近邻样本的度量的有效性。不同增强样本下的实验结果如图 3 所示,其中 TopoSMOTE-edge 代表不考虑具有边连接的少数类样本选取,而只在节点中进行采样,TopoSMOTE-low 代表选择最近邻样本时不选择使用低维映射空间,而使用原始的高维空间计算距离。

从图 3 所示对比结果可以看出,TopoSMOTE 在两种简化条件下的性能均有所下降,这验证了考虑拓扑关系和使用低维空间计算距离在提升检测性能中的有效性。

4 结论

通过本文的研究和实验,得到了以下结论。

(1)TopoSMOTE 方法通过在拓扑空间中生成新的少数类样本,可以有效提升少数类攻击类型的特征质量,有助于更准确地识别网络流量中的异常行为有效地处理了类别不平衡问题。

(2)TopoSMOTE 采用了一种新的样本选择策略,通过低维空间中的距离度量选择具有拓扑关系的最近邻样本进行合成。不仅保持了原始数据的拓扑结构,还有效避免了传统过采样方法中可能产生的过拟合和生成域外样本的问题。

(3)实验结果表明,TopoSMOTE 在 CIC-IDS2017 和 CIC-MalMen2022 数据集上的整体性能优于其他方法,验证了 TopoSMOTE 方法在应对类不平衡问题上的有效性。

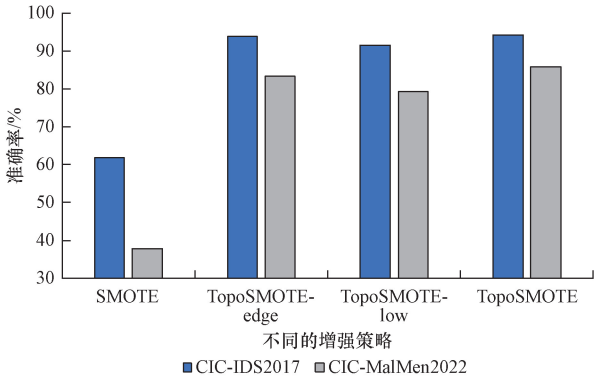

图 3 不同增强样本下的实验结果

Fig. 3 Experimental results under different enhanced samples

表 6 不同模型在 CIC-IDS2017 数据集上的宏观指标比较

Table 6 Comparison of macro indicators of different models on CIC-IDS2017 dataset

检测模型	*A*/%	*P*/%	*R*/%	FPR/%	F_1/%
DNN	87.23	61.84	61.31	1.13	61.84
AB-LightGBM	96.28	83.04	83.50	0.86	83.27
SMOTE-TomekLink	94.96	81.12	82.45	1.23	81.78
TMG-IDS	95.53	86.86	89.69	0.87	88.25
TopoSMOTE	96.12	93.17	95.34	0.49	94.25

表 7 不同模型在 CIC-MalMen2022 数据集上的宏观指标比较

Table 7 Comparison of macro indicators of different models on CIC-MalMen2022 dataset

检测模型	*A*/%	*P*/%	*R*/%	FPR/%	F_1/%
DNN	65.74	40.08	34.50	2.18	32.08
AB-LightGBM	89.99	78.94	67.13	1.09	71.02
SMOTE-TomekLink	83.94	58.51	61.24	1.48	59.08
TMG-IDS	85.65	74.77	74.96	1.12	74.86
TopoSMOTE	93.41	85.44	86.37	1.03	85.90

参 考 文 献

[1] Buczak A L, Guven E. A survey of data mining and machine learning methods for cyber security intrusion detection[J]. IEEE Communications Surveys & Tutorials, 2016, 18(2): 1153-1176.

[2] Butun I, Morgera S D, Sankar R. A survey of intrusion detection systems in wireless sensor networks[J]. IEEE Communications Surveys & Tutorials, 2013, 16(1): 266-282.

[3] Das S, Saha S, Priyoti A T, et al. Network intrusion detection and comparative analysis using ensemble machine learning and feature selection[J]. IEEE Transactions on Network and Service Management, 2021, 19(4): 4821-4833.

[4] 杨锦溦, 杨宇, 姚铖鹏, 等. 基于改进深度卷积生成对抗网络的入侵检测方法[J]. 科学技术与工程, 2022, 22(8): 3209-3215.
Yang Jinwei, Yang Yu, Yao Chengpeng, et al. Intrusion detection method based on improved deep convolutional generative adversarial network[J]. Science Technology and Engineering, 2022, 22(8): 3209-3215.

[5] Abdel-Basset M, Chang V, Hawash H, et al. Deep-IFS: intrusion detection approach for industrial internet of things traffic in fog environment[J]. IEEE Transactions on Industrial Informatics, 2020, 17(11): 7704-7715.

[6] Wheelus C, Bou-Harb E, Zhu X. Tackling class imbalance in cyber security datasets[C]//2018 IEEE International Conference on Information Reuse and Integration (IRI). New York: IEEE, 2018: 229-232.

[7] 孙佳佳, 李承礼, 常德显, 等. 基于生成对抗网络的入侵检测类别不平衡问题数据增强方法[J]. 科学技术与工程, 2022, 22(18): 7965-7971.
Sun Jiajia, Li Chengli, Chang Dexian, et al. Data augmentation method for intrusion detection imbalance problem using generative adversarial networks[J]. Science Technology and Engineering, 2022, 22(18): 7965-7971.

[8] 武洋名, 宗学军, 何戡. 基于数据增强的 DBN-ELM 入侵检测方法[J]. 科学技术与工程, 2022, 22(34): 15195-15202.
Wu Yangming, Zong Xuejun, He Kan. DBN-ELM intrusion detection method based on data augmentation[J]. Science Technology and Engineering, 2022, 22(34): 15195-15202.

[9] 郭文婷, 张军, 魏洪伟, 等. 基于欠采样和对抗自编码器的入侵检测算法[J]. 信息通信, 2019, 12: 58-60.
Guo Wenting, Zhang Jun, Wei Hongwei, et al. An intrusion detection algorithm based on undersampling and adversarial autoencoder[J]. Information & Communications, 2019, 12: 58-60.

[10] Zhao T, Zhang X, Wang S. Graphsmote: imbalanced node classification on graphs with graph neural networks[C]//Proceedings of the 14th ACM International Conference on Web Search and Data Mining. New York: ACM, 2021: 833-841.

[11] 周杰英, 贺鹏飞, 邱荣发, 等. 融合随机森林和梯度提升树的入侵检测研究[J]. 软件学报, 2021, 32(10): 3254-3265.
Zhou Jieying, He Pengfei, Qiu Rongfa, et al. Research on intrusion detection based on random forest and gradient boosting tree[J]. Journal of Software, 2021, 32(10): 3254-3265.

[12] Ren H, Tang Y, Dong W, et al. DUEN: dynamic ensemble handling class imbalance in network intrusion detection[J]. Expert Systems with Applications, 2023, 229: 120420.

[13] Telikani A, Shen J, Yang J, et al. Industrial IoT intrusion detection via evolutionary cost-sensitive learning and fog computing[J]. IEEE Internet of Things Journal, 2022, 9(22): 23260-23271.

[14] Zhou Z H, Liu X Y. Training cost-sensitive neural networks with methods addressing the class imbalance problem[J]. IEEE Transactions on Knowledge and Data Engineering, 2005, 18(1): 63-77.

[15] Chen Y, Ashizawa N, Yean S, et al. Self-organizing map assisted deep autoencoding gaussian mixture model for intrusion detection[C]//2021 IEEE 18th Annual Consumer Communications & Networking Conference (CCNC). New York: IEEE, 2021: 1-6.

[16] Zia A, Khamis A, Nichols J, et al. Topological deep learning: a review of an emerging paradigm[J]. Artificial Intelligence Review, 2024, 57(4): 77.

[17] Wheelus C, Bou-Harb E, Zhu X. Tackling class imbalance in cyber security datasets[C]//2018 IEEE International Conference on Information Reuse and Integration (IRI). New York: IEEE, 2018: 229-232.

[18] Sun Z, Song Q, Zhu X, et al. A novel ensemble method for classifying imbalanced data[J]. Pattern Recognition, 2015, 48(5): 1623-1637.

[19] Zhang Y, Liu Q. On IoT intrusion detection based on data augmentation for enhancing learning on unbalanced samples[J]. Future Generation Computer Systems, 2022, 133: 213-227.

[20] Gong Z, Jiang J, Jiang N, et al. An improved hybrid sampling model for network intrusion detection based on data imbalance[C]//International Conference on Artificial Intelligence Security and Privacy. Singapore: Springer Nature Singapore, 2023: 164-175.

[21] Bedi P, Gupta N, Jindal V. Siam-IDS: Handling class imbalance problem in intrusion detection systems using siamese neural network[J]. Procedia Computer Science, 2020, 171: 780-789.

[22] 潘桐, 陈伟, 吴礼发. 面向不平衡样本的物联网入侵检测方法[J]. 网络与信息安全学报, 2023, 9(1): 130-139.
Pan Tong, Chen Wei, Wu Lifa. IoT intrusion detection method for unbalanced samples[J]. Chinese Journal of Network and Information Security, 2023, 9(1): 130-139.

[23] Ding H, Chen L, Dong L, et al. Imbalanced data classification: a KNN and generative adversarial networks-based hybrid approach for intrusion detection[J]. Future Generation Computer Systems, 2022, 131: 240-254.

[24] Singh G, Mémoli F, Carlsson G E. Topological methods for the analysis of high dimensional data sets and 3d object recognition[J]. Eurographics Symposium on Point-Based Graphics, 2007, 2: 91-100.

[25] Mi H, Huang X, Muruganujan A, et al. PANTHER version 11: expanded annotation data from Gene Ontology and Reactome pathways, and data analysis tool enhancements[J]. Nucleic Acids Research, 2017, 45(D1): D183-D189.

[26] Skaf Y, Laubenbacher R. Topological data analysis in biomedicine: a review[J]. Journal of Biomedical Informatics, 2022, 130: 104082.

[27] Altındiş F, Yılmaz B, Borisenok S, et al. Parameter investigation of topological data analysis for EEG signals[J]. Biomedical Signal Processing and Control, 2021, 63: 102196.

[28] Giansiracusa N, Giansiracusa R, Moon C. Persistent homology machine learning for fingerprint classification[C]//2019 18th IEEE International Conference on Machine Learning and Applications (ICMLA). New York: IEEE, 2019: 1219-1226.

[29] Songdechakraiwut T, Chung M K. Dynamic topological data analysis for functional brain signals[C]//2020 IEEE 17th International Symposium on Biomedical Imaging Workshops (ISBI Workshops). New York: IEEE, 2020: 1-4.

[30] Jeon E S, Choi H, Shukla A, et al. Topological knowledge distillation for wearable sensor data[C]//2022 56th Asilomar Conference on Signals, Systems, and Computers. New York: IEEE, 2022: 837-842.

[31] Monkam G F, De Lucia M J, Bastian N D. A topological data analysis approach for detecting data poisoning attacks against machine learning based network intrusion detection systems[J]. Computers & Security, 2024(9): 103929.

[32] McInnes L, Healy J, Saul N, et al. UMAP: uniform manifold approximation and projection[J]. Journal of Open Source Software, 2018, 29(3): 861.

[33] Liu Z, Hu C, Shan C. Riemannian manifold on stream data: fourier transform and entropy-based DDoS attacks detection method [J]. Computers & Security, 2021, 109: 102392.

[34] Han H, Wang W Y, Mao B H. Borderline-SMOTE: a new oversampling method in imbalanced data sets learning[C]//International Conference on Intelligent Computing. Berlin: Springer Berlin Heidelberg, 2005: 878-887.

[35] Sharafaldin I, Lashkari A H, Ghorbani A A. Toward generating a new intrusion detection dataset and intrusion traffic characterization [C]//Proceedings of the 8th International Conference on Information Systems Security and Privacy (ICISSP). Lisboa: Science and Technology Publications, 2018: 108-116.

[36] Carrier T, Victor P, Tekeoglu A, et al. Detecting obfuscated malware using memory feature engineering[C]//Proceedings of the 8th International Conference on Information Systems Security and Privacy (ICISSP). Lisboa: Science and Technology Publications, 2022: 177-188.

[37] Chen W, Wang H, Fei M, et al. An intrusion detection method using adasyn and bayesian optimized lightgbm[C]//2022 34th Chinese Control and Decision Conference (CCDC). New York: IEEE, 2022: 4622-4627.

[38] Talukder M A, Sharmin S, Uddin M A, et al. MLSTL-WSN: machine learning-based intrusion detection using SMOTETomek in WSNs[J]. International Journal of Information Security, 2024, 23(3): 2139-2158.

[39] Ding H, Sun Y, Huang N, et al. TMG-GAN: Generative adversarial networks-based imbalanced learning for network intrusion detection[J]. IEEE Transactions on Information Forensics and Security, 2023, 19: 1156-1167.