计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2022,31(10):51−59 [doi: 10.15888/j.cnki.csa.008730]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于集成 SVM 和 Bagging 的未知恶意流量检测①

赵　静[1,2], 李　俊[1,2], 龙　春[1,2], 杜冠瑶[1,2], 万　巍[1,2], 魏金侠[1]

[1](中国科学院 计算机网络信息中心, 北京 100190)
[2](中国科学院大学, 北京 100049)
通信作者: 龙　春, E-mail: anquanip@cnic.cn

摘　要: 未知恶意网络流量检测是异常检测领域亟待解决的核心问题之一. 从高速网络数据流中获取的流量数据往往具有不平衡性和多变性. 虽然在恶意网络流量异常检测特征处理和检测方法方面已存在诸多研究, 但这些方法在同时解决数据不平衡性和多变性以及模型检测性能方面仍存在不足. 因此, 本文针对未知恶意网络流量检测目前存在的困难, 提出了一种基于集成 SVM 和 Bagging 的未知恶意流量检测模型. 首先, 针对网络流量数据的不平衡性, 提出一种基于 Multi-SMOTE 过采样的流量处理方法, 以提高流量处理后的特征质量; 第二, 针对网络流量数据分布的多样性, 提出一种基于半监督谱聚类的未知流量筛选方法, 以实现从具有多样分布的混合流量中筛选出未知流量; 最后, 基于 Bagging 思想, 训练了集成 SVM 未知恶意流量检测器. 实验结果表明, 本文所提出的基于集成 SVM 与 Bagging 的未知流量攻击类型检测模型在综合评价 ($F1$ 分值) 上优于目前同类未知恶意流量检测方法, 同时在不同数据集上具有较好的泛化能力.

关键词: 未知恶意流量检测; Multi-SMOTE 过采样; 半监督谱聚类; 集成学习; 支持向量机

引用格式: 赵静,李俊,龙春,杜冠瑶,万巍,魏金侠.基于集成 SVM 和 Bagging 的未知恶意流量检测.计算机系统应用,2022,31(10):51–59. http://www.c-s-a.org.cn/1003-3254/8730.html

Unknown Malicious Traffic Detection Based on Integrated SVM and Bagging

ZHAO Jing[1,2], LI Jun[1,2], LONG Chun[1,2], DU Guan-Yao[1,2], WAN Wei[1,2], WEI Jin-Xia[1]

[1](Computer Network Information Center, Chinese Academy of Sciences, Beijing 100190, China)
[2](University of Chinese Academy of Sciences, Beijing 100049, China)

Abstract: Unknown malicious network traffic detection is one of the core problems to be solved in anomaly detection as the traffic data obtained from high-speed network data flow are often unbalanced and changeable. Although there have been many studies on feature processing and detection methods of unknown malicious network traffic detection, these methods have shortcomings in simultaneously solving data imbalance and variability as well as detection performance. Considering the difficulty in unknown malicious network traffic detection, this study proposes an unknown malicious traffic detection model based on integrated SVM and bagging. Firstly, in view of the imbalance of network traffic data, a traffic processing method based on Multi-SMOTE oversampling is put forward to improve the feature quality upon traffic processing. Secondly, considering the distribution diversity of network traffic data, an unknown traffic screening method based on semi-supervised spectral clustering is presented to screen unknown traffic from mixed traffic with a diverse distribution. Finally, with the idea of Bagging, an unknown malicious traffic detector based on integrated SVM is trained. The experimental results reveal that the proposed detection model is superior to the current similar methods in comprehensive evaluation ($F1$ value), and it also has good generalization ability on different data sets.

Key words: unknown malicious traffic detection; Multi-SMOTE oversampling; semi-supervised spectral clustering; Bagging; support vector machine (SVM)

① 基金项目: 国家自然科学基金 (61672490)
收稿时间: 2022-01-04; 修改时间: 2022-01-29; 采用时间: 2022-02-22; csa 在线出版时间: 2022-06-24

网络空间是人们生产、生活的重要空间, 网络安全已经成为国家安全的重要组成部分, 我国是遭受网络攻击最严重的国家之一, 重大网络安全事件时有发生. 仅在 2021 年上半年, 我国约有 446 万台主机感染恶意程序[1], 平均每月约有 4 300 起峰值超过 10 Gb/s 大流量 DDoS 攻击, 并且总体呈现递增趋势. 基于流量的攻击增长如此迅速的主要原因是新型未知恶意流量的兴起与发展.

僵尸网络、勒索病毒、网络蠕虫木马、拒绝服务攻击等网络恶意技术手段更加多样化、恶意代码变异更加快速. 恶意攻击在网络环境中具有高度复杂性和多样性, 对新型未知的网络恶意行为进行检测成为异常检测技术研究的重要新方向[2–4].

研究人员对大量恶意数据流样本分析发现, 大部分新出现的恶意数据流实际上是已知恶意流量的变体形式, 具有一定的关联性[5]. 恶意流量的执行者通过多态、变形等技术手段打乱已知恶意流量的特征, 形成新的恶意流量变种, 以绕过安全设备的检测. 为了能够快速有效地对抗新型安全威胁, 建立高效的分类方法对海量未知样本进行检测是有必要的.

在传统已有的异常检测方法中, 专门对未知恶意流量异常检测特征处理和检测方法进行研究的成果并不是很多, 并且在同时解决数据不平衡性、多变性和模型检测准确性等方面存在不足. 具体表现在以下方面.

当前网络流量攻击手段变得愈发隐蔽和复杂, 关键攻击特征常常隐藏在正常流量数据流中, 攻击样本数量极少, 同时复杂网络流量攻击数据特征维度比较高, 在这种攻击样本数量严重不均衡且特征维度过高的情况下传统模型训练并不完善, 缺乏针对不均衡性、特征维度过高的恶意流量进行特征全面提取的有效手段; 传统的恶意流量检测方法未能充分考虑样本分布, 大多仅考虑凸样本空间分布, 缺乏对样本分布适应性较强的未知流量准确识别方法; 同时传统恶意流量检测方法也面临对新输入样本的检测需要更新整个模型参数的问题, 对恶意流量变化快速更新的能力不足, 模型的实时性较低.

综上, 当前形势下传统的网络未知恶意流量检测技术已经面临严峻的挑战. 针对真实网络环境中恶意流量数据极度不平衡的情况下, 研究具有高检测率、低误报率的未知恶意流量检测方法是网络空间安全领域的迫切需求.

因此, 本文综合考虑传统恶意流量检测方法中样本不均衡性、样本分布适应性不足等问题, 采用机器学习方法对未知恶意流量高效检测进行深入研究. 本文的主要贡献总结如下:

首先, 针对网络流量数据不平衡问题, 提出基于 Multi-SMOTE 过采样的流量处理方法, 该方法可以为后续未知攻击检测步骤提供高质量的数据集, 降低数据质量导致的检测误差.

第二, 提出基于半监督谱聚类的未知流量筛选方法. 由于未知流量往往掺杂在海量已知流量中, 对未知流量先进行筛选才能使后续恶意未知流量检测更准确、快速和便捷. 而网络数据流特征复杂、属性维度过高, 利用传统的检测方法或者聚类算法识别未知流量往往在运行时间和准确率等方面不具有优势. 鉴于此, 考虑利用一种对数据分布适应性更强的方法来实现对未知流量的精确识别. 谱聚类具有识别非凸分布聚类的能力, 能在任意形状的样本空间上聚类, 因此, 提出一种基于半监督谱聚类的未知流量筛选方法, 能够从混合流量中精准识别未知样本.

第三, 利用 Bagging 思想, 训练基于集成 SVM 的未知恶意流量检测器. 在每次训练时, 上一轮被分错的样本在本轮以 Bagging 采样的方式选出来加入训练集中, 进行反复迭代与训练, 得到最优的参数, 最终获得综合性能高的检测器.

1 相关研究与分析

下面分别从本文应用到的过采样技术、未知流量攻击类型检测两个方面论述研究现状及发展动态, 其中未知流量类型检测包括未知流量筛选和未知流量攻击类型检测两部分.

1.1 过采样技术

过采样技术可以有效解决因样本不均衡导致的机器学习模型因对样本量较小的数据无法充分学习到特征, 因而造成欠拟合的问题. 2002 年, 文献 [6] 提出了 SMOTE 过采样算法, 该算法是对随机过采样方法的改进, 是对每个少数类样本, 从它的最近邻中随机选择一个样本, 然后在原样本和被选样本之间的连线上, 随机选择一点作为新合成的少数类样本. SMOTE 算法摒弃了随机过采样复制样本的方法, 可以防止随机过采用中容易出现的过拟合问题. 文献 [7] 中作者认为有些样本远离边界, 对分类没有多大帮助, 将少数类样本根据

与多数样本的距离大小分为 Noise, Safe, Danger 三类样本集, 然后只对 Danger 中的样本集合使用 SMOTE 算法. 针对 SMOTE 对每个少数样本合成相同数量的样本, 文献 [8] 提出了自适应合成抽样算法 ADASYN, 可以采用某种机制自动决定每个少数样本产生样本的数量, 以保证数据分布不发生过大变化. 文献 [9] 通过考虑数据集中确定性的变化来添加新的数据点, 根据权重确定少数样本被选为种子的概率. 文献 [10] 提出了一种新的基于核的自适应合成过采样方法, 称为 Kernel-ADASYN, 用于不平衡数据分类问题. 其思想是构造一个自适应过采样分布来生成合成的少数群体数据. 首先用核密度估计方法估计自适应过采样分布, 然后根据不同少数群体数据的难度对自适应过采样分布进行加权. 文献 [11] 针对多类别不平衡问题, 提出 SMOM 算法, 通过对辅助样本的选择确定合成样本的位置. 文献 [12] 针对传统采样方式的准确率和鲁棒性欠佳, 容易丢失重要样本信息的问题, 提出了一种基于样本特征的自适应变邻域 SMOTE 算法, 实验表明该方法比其他传统方法有更好的准确率和鲁棒性. 文献 [13] 认为 SMOTE 算法没有涉及缺失值的恢复, 虽然 FID 算法[14] 解决了这个问题, 但没有很好地考虑到数据属性之间的相关性, 为此提出了一种基于模糊规则的过采样技术, 高效地解决了数据不平衡和缺失数据这两个问题. 文献 [15] 在传统过采样 SMOTE 算法的基础上, 提出了 LR-SMOTE 算法, 结合 K-means 和 SVM 方法使新生成的样本靠近样本中心, 避免产生离群样本或改变数据集的分布.

然而, 现有的 SMOTE 及其改进算法大多关注合成数据的质量, 而忽视了数据分布和数据噪声问题. 现实中数据是以流的方式出现, 数据分布随时间不断变化, 容易导致结构不稳定而产生概念漂移, 因此在过采样过程中, 会加剧少量样本数据噪声的叠加.

1.2 未知恶意流量检测

在未知恶意流量检测方面, 近年来, 为了躲避检测软件的查杀, 恶意流量在传播过程中采取多态[16] 等变种技术伪装成与已知攻击流量不同的形式去传播. 因此, 对基于已知攻击流量伪装成新型攻击流量的检测问题也变得越来越重要[17–24].

文献 [25] 将决策树和神经网络进行结合, 采用决策树提取规则, 提高了算法的精确度. 文献 [26] 介绍了一种因果决策树模型, 中间节点具有因果解释的作用, 且因果决策树算法具有可扩展性, 分类算法在保持分类精度较高的前提下还可以显著降低算法的执行复杂度, 从整体上提升异常检测模型的性能.

文献 [27] 提出一种多分类器融合的检测模型, 该模型克服了传统异常检测系统存在的虚警率高、实时性好、可扩展性差等问题, 是一种增量式机器学习分类器, 用于对网络数据流进行智能检测和分析. 该模型虽然具有很好的实时性, 但是在恶意流量多态或者演变形式的检测方面的检测率不是很高. 文献 [28] 提出了一种基于异常的异常检测系统, 该系统使用集成分类方法来检测 Web 服务器上的未知攻击. 该过程涉及利用过滤器和包装器选择程序去除不相关和多余的特征. 然后将 Logitboost 与随机森林一起用作弱分类器. 文献 [29] 提出了两个基于蚁群算法 (ACA) 的未知攻击的异常检测系统. 该 IDS 可以在未标记的数据集上学习并检测未知攻击. 其提出的 IDS 是 ACA 和其他监督学习算法的结合, 并将决策树 (DT) 和人工神经网络 (ANN) 分别与 ACA 相结合, 虽然取得较高的检测率, 但是误报率并不理想.

文献 [30] 提出了一种基于信息增益率和半监督学习的异常检测方案, 针对未知攻击网络流量特征难以定量选取、动态变化的攻击难以自适应地应对, 以及训练数据集规模过小而导致模型难以训练 3 个问题, 采用半监督学习算法通过少量已标记数据生成大规模训练数据集, 以此对检测模型进行训练, 并引入信息增益率对网络流量特征自适应地定量提取以实现对目标网络中未知攻击的检测.

支持向量机 (SVM) 是应用于异常检测领域最广泛的一种分类方法[31,32]. 文献 [33] 提出了一种基于遗传算法和支持向量机的异常检测方法. 该模型采用遗传算法进行特征选择, 并在适应度函数上进行了创新, 有效地减少了数据的规模和维数, 从而显著减少模型训练的时间, 并同时达到较高的准确率. 文献 [34] 提出一种基于 SVM 的异常检测与时间混沌粒子群优化 (TVCPSO) 方法的结合模型, 该模型利用 TVCPSO 确定 SVM 分类器的最佳参数, 在检测新型未知恶意流量方面具有较强的泛化能力.

文献 [35] 分析当前流量攻击检测工作研究现状与面临的挑战, 指出目前大多数的异常检测系统在误报率和检测率方面的效果不是很理想, 为了解决这些问题, 多数学者重点研究集成分类器的研发, 通过对多个

但分类器进行组合, 获得集成分类器. 集成分类器可以避免单个分类器的不足, 增强分类器的整体性能.

综上所述, 本文综合考虑传统恶意流量检测方法中对样本不均衡性与维度过高处理不完善、样本分布适应性不足、检测模型更新能力较弱等问题, 采用机器学习方法及集成分类器对未知恶意流量高效检测进行深入研究.

2 本文方法

本节将详细介绍基于集成 SVM 与 Bagging 的未知恶意流量检测模型的设计. 首先提出基于 Multi-SMOTE 过采样的流量处理方法, 以解决实际网络流量数据中异常样本的极度不平衡性导致的检测模型训练欠拟合问题, 为后续检测步骤提供高质量的数据集. 第二, 针对网络流量中数据分布的多样性, 提出基于半监督谱聚类的未知流量筛选方法, 能够从具有多样分布的混合流量中筛选出未知流量, 以便后续步骤实现未知恶意流量的分类. 最后, 利用 Bagging 思路, 经过多轮调参, 训练了一个能够识别未知恶意流量的集成 SVM 分类器. 实验结果表明, 本文提出的模型针对未知流量恶意识别具有较好的综合性能.

整体的方法流程图如图 1 所示.

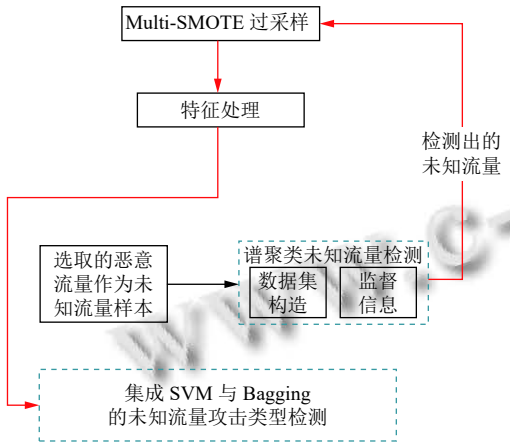

图 1 基于集成 SVM 与 Bagging 的未知流量攻击类型检测模型整体流程

2.1 基于 Multi-SMOTE 的过采样方法

通常, 网络流量数据中会存在某些新型恶意行为的数目相比其他恶意行为类别少很多, 出现类别数量极端不平衡的现象. 直接用这些数据集训练检测模型, 很容易引起过拟合. 为了避免这种现象发生, 首先对少数类的恶意行为样本进行过采样, 避免不同恶意行为数量不平衡现象的发生.

本文基于 SMOTE 和 Borderline-SMOTE 提出一种改进的过采样算法, 用于扩展新样本的生成区域, 方法命名为 Multi-SMOTE. 假设少数类样本集为X, 多数类样本集为N, 且$X = \{x_1, x_2, \cdots, x_{xnum}\}$, $N = \{n_1, n_2, \cdots, n_{nnum}\}$, 其中$xnum$和$nnum$分别为$X$和$N$的样本数目, 每个样本有$d$个特征, 则每一个少数类样本可表示为$x$: $(x^1, x^2, \cdots, x^d)^{\mathrm{T}}$. 首先, 利用 Borderline-SMOTE 的关键少数类样本选取算法来产生关键样本集key_set. 基于上述方法得到的key_set, 针对该集合中的每一个样本进行过采样操作, 可以有效克服原始 SMOTE 产生过多无用样本点的缺陷. 为解决 SMOTE 中存在的过采样区域过于狭窄的问题, 针对key_set中的每一个样本点应用 Multi-SMOTE 算法. 利用 Borderline-SMOTE 算法得到key_set之后, 针对key_set中的每一个样本x, 找到x在X中的K个近邻, 并同时找到这K个近邻中与x之间欧氏距离最大的近邻, 用x_T表示. 以x为中心, $\|x - x_T\|_2$为半径 (用r表示), 在高维空间中确定一个区域, 新样本将在该区域中合生, 该区域用G表示.

由于已将生成区域固定, 因此可以得到新样本每一维特征的取值范围.针对第l个特征, $l = 1, 2, \cdots, d$, 其取值范围为: $[x^l - r, x^l + r]$. 之后, 从均匀分布$(-1, +1)$中生成d个随机数, 将其表示为σ_l. 使用σ_l可直接合成一个新样本$\tilde{x}$, $\tilde{x} = (x^1 + \sigma_1 r, x^2 + \sigma_2 r, \cdots, x^d + \sigma_d r)$. 然而, 即使每一维特征的取值均在规定的取值范围内, 仍然无法保证该样本在G当中, 因为G是一个超球体区域, 而各特征的取值范围确定的是超方体区域, 此时$\tilde{x}$与x之间的欧氏距离无法保证为r.

为了解决上述问题, 对σ_l进行归一化操作, 令$\tilde{\sigma}_l = \sigma_l / \sqrt{m}$, m是σ_l^2的加和. 基于$\tilde{\sigma}_l$合成一个新样本$\tilde{x}$, $\tilde{x} = (x^1 + \tilde{\sigma}_1 r, x^2 + \tilde{\sigma}_2 r, \cdots, x^d + \tilde{\sigma}_d r)$, $\tilde{x}$在G的边界之上. 基于$\tilde{x}$, 便可以在G中生成一个新的样本x_{new}, $x_{\mathrm{new}} = x + \rho(\tilde{x} - x)$, $\rho \in (0, 1)$.

2.2 基于半监督谱聚类的未知流量筛选

网络流量数据来源广、层次多、差异大、维度高、内在关系错综复杂, 未知恶意流量隐藏得比较深, 为了能更准确地对未知流量的攻击类型进行识别, 首先将未知流量和已知流量区分开来. 因此, 首先解决的是未知流量的筛选问题, 考虑谱聚类方法对数据分布的适应性更强, 具有识别非凸分布聚类的能力, 能在任

意形状的样本空间上聚类, 且收敛于全局最优解, 基于此本文建立一种基于半监督谱聚类的未知流量筛选模型.

2.2.1 数据集生成

选定一组已知网络流量的数据集S, 去掉数据集S中的标签信息, 形成一组无标签数据集S'. 将其与模拟生成的未知流量混合到一起, 形成一个无标签的综合数据集M, 作为谱聚类的训练数据集. 本文选择以去掉有标签数据标签的方式来形成无标签的混合数据集, 除了构造监督信息, 还可以通过模型最后的分类结果与标签对比来验证半监督谱聚类的性能.

2.2.2 监督信息构造

半监督聚类中, 监督信息能有效改善聚类算法的性能. 有国外学者证实, 寻找满足所有监督信息的聚类解是一个 NP 完备问题, 监督信息越多, 半监督聚类算法的复杂度越高, 但聚类性能不一定越高, 因此挖掘适合半监督聚类的监督信息十分关键.

本文为了调整谱聚类算法距离矩阵的元素值, 构造成对的监督信息. 通常, 在聚类过程中, 距离比较远的数据被认为是属于不同类的, 从而被划分到不同的类中. 同样, 距离较近的数据被认为是属于相同类的, 从而被划分到同一类别中. 因此, 本文将上述数据标记为两类集合作为监督信息, 分别表示距离远的同类集X和距离近的不同类集C.

2.2.3 半监督谱聚类

基于以上生成的数据集和监督信息, 提出了半监督谱聚类算法如算法 1 所示.

算法 1. 基于半监督谱聚类的未知流量筛选算法

输入: 已知网络流量数据集 , 无标签的综合训练数据集M, 距离矩阵D
输出: 未知流量类集

1) 为距离矩阵D中元素赋初值 0, 计算数据集中两点之间欧氏距离;
2) 修改距离矩阵D, 若这两点属于同类集X, 则矩阵元素为 0; 若这两点属于不同类集C, 则矩阵元素为无穷;
3) 构造矩阵S, 其各个元素为距离矩阵的倒数;
4) 构造矩阵$P=L^{-1/2}SL^{1/2}$, 其中L为对角矩阵$L_{ii}=\sum_{j=1}^{n}S_{ij}$;
5) 经过谱聚类过程获得 2 个类;
6) 对已知流量的数据集进行聚类, 分别计算上述 2 个类的聚类中心与已知数据集中每个类聚类中心的平均距离, 比较 2 个类到已知数据集的平均距离大小.

2.3 基于集成 SVM 和 Bagging 的恶意流量判定方法

网络流量数据量大、维度高, 内部关系复杂, 应用传统的统计方法不能高效率分析和处理. 支持向量机不同于现有的统计分析方法, 避开了从归纳到演绎的传统过程, 可以高效地实现从训练样本到预测样本的转导推理, 大大简化了分类与回归问题, 同时具有很好的鲁棒性.

集成学习通过将多个学习器进行结合, 可获得比单一学习器更见显著的泛化性能. 本文基于 AdaBoost 和 Bagging 结合的方法实现对未知流量攻击类型的检测, 根据集成过程中弱分类器的权重调整训练样本的数据量, 依据 Bagging 采样的方式选取训练样本. 最终的改进分类器$F(\cdot)$是根据基分类器的加权求和获得的. 本文研究目的是实现未知流量攻击类型的检测, 因此, 集成 SVM 分类器指的是多分类器, 涉及到的 SVM 基分类器是适用于多分类的场景.

对于 SVM, 在集成过程中调整样本的权重只会改变样本在空间中的位置, 并且不会降低分类错误的样本的损失. 其实, 在上一轮训练过程中被分类错误的样本应在本轮训练中会带来更大的损失, 这样的话需要选择较大的惩罚参数C来平衡分错样本带来的损失. 但是, 为不同的样本选择不同的惩罚参数C比较困难. 因此, 利用 Bagging 的采样思想来选择每一轮训练的样本, 以提高精度和召回率异常检测模型.

将上一轮分错的样本复制α ($\alpha>1$)份, 复制之后的所有样本将被加入到本轮的训练样本中. 显然, 在本轮的训练样本中, 被分错的样本数量增加了. 因此, 本轮训练中, 分类器倾向于将上轮分错的样本以更高的准确率正确分类. 具体方法描述如下:

令M为所有训练样本的集合, N是本轮训练中选择的训练样本, $N<|M|$. 假设本轮训练过程中分类错误的样本表示为集合Q, 则从M中随机选择$N-\alpha|Q|$个样本形成样本集合P. 将集合Q中的样本复制α次, 复制之后的样本的集合表示为Q^{α}. 最终集合P和集合Q^{α}组合成为下一轮训练过程中的训练数据集. 为了避免离群点对训练样本的重采样造成影响, 设定错位分类阈值的上限H和下限L. 具体算法如算法 2 所示.

3 实验及结果分析

3.1 实验环境及数据说明

本文实验环境为 Windows 7 平台, Intel Core i7, 8 核 CPU, 分配内存 20 GB. 实验采用 Python 语言作为编程语言. 本文定义模型性能评估指标分别为准确率 (accuracy, ACC)、精确率 (precision)、召回率 (recall)、误报率 (false positive rate, FPR) 和 $F1$ 分值 ($F1$-score).

算法 2. 基于集成 SVM 和 Bagging 的恶意流量判定方法

输入: 分类错误样本集Q, 复制之后的样本集Q^{α}, 模型参数N、α、L、H
输出: 改进分类器$F(\cdot)$, 测试数据的具体攻击类型

1) 如果$|Q|<L$, 从训练数据集中随机选择N个样本;
2) 如果$|Q|>H$, 从集合Q中随机选择H个样本形成新的集合Q, 然后从训练样本集合M中随机选择$N-\alpha|Q|$个样本形成集合P, P和Q^{α}组合作为下一轮训练过程的训练样本, 该过程与 AdaBoost 权重调整策略类似;
3) 如果$L\leqslant|Q|\leqslant H$, 从训练数据集中随机选取$N-\alpha|Q|$个样本形成集合P, P和Q^{α}组合作为下一轮训练过程的训练样本;
4) 以这样的方式生成多个基分类器, 然后根据基分类器的加权求和获得改进分类器$F(\cdot)$, N, α, L和H是超参数, 通过十折交叉验证的方式获得最优值;
5) 将测试数据集输入$F(\cdot)$中, 输出结果显示测试数据的具体攻击类型.

数据集方面, 首先, 本文采用流量生成工具 Flightsim 自己生成流量数据集S, 以测试 Multi-SMOTE 过采样算法、半监督谱聚类算法和未知流量检测模型性能. Flightsim 是一款轻量级的开源网络安全工具, 安全研究人员可以利用这款工具来生成恶意网络流量. 其次, 为了进一步评估检测模型的性能, 本文与其他同类论文中的方法进行了对比实验, 为了使实验具有可比性, 采用了同类论文中所使用的 UCI 数据集和 KDD'99 数据集. 最后, 为了验证本文所提方法的泛化能力, 分别在自生成数据集S、UCI 数据集和 KDD'99 数据集进行了实验.

3.2 Multi-SMOTE 过采样算法性能评估

本文利用 Flightsim 流量生成工具, 模拟生成了 9 类流量数据, 分为 1 类正常流量数据和 8 类恶意流量数据, 自生成流量数据集的数据分布如图 2 所示.

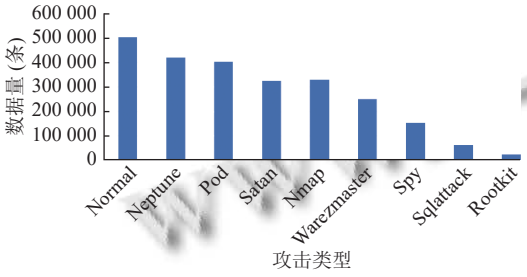

图 2 自生成流量数据集数据分布图

由图 2 可以看出, 恶意流量数据存在着严重的数据不平衡问题, sqlattack 和 rootkit 恶意流量占比远远小于 normal 类. 针对这一问题, 本文用提出的 Multi-SMOTE 过采样方法对数据进行采样, 采样后的数据分布如图 3 所示.

经过 Multi-SMOTE 采样后, 少量样本数据有了大幅度增加, 并且在数据量增加的同时保证了数据分布没有发生太大变化. 本文提出的 Multi-SMOTE 方法解决了先前过采样算法存在的数据分布不平衡和数据噪声叠加的问题. 在数据预处理阶段便获得质量较好的训练数据, 有利于训练出更好的模型.

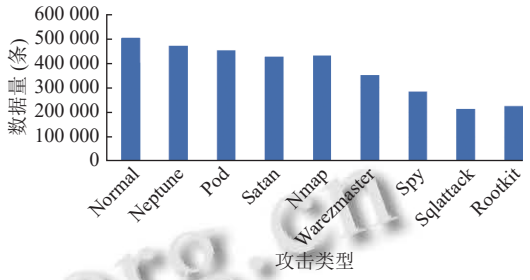

图 3 Multi-SMOTE 过采样后数据分布图

3.3 半监督谱聚类算法性能评估

为了在自生成数据集上测试本文提出的半监督谱聚类算法, 我们将随机从自生成数据集中抽取 25% 的样本作为子数据集进行聚类实验. 将子数据集S上的所有标签信息去掉, 形成一组无标签数据集S'. 将S'输入到半监督谱聚类算法中进行聚类实验. 为了验证本文提出的半监督谱聚类算法的有效性, 将相同的数据采用 K-means 算法进行对比实验.

从表 1 可以看出, 本文提出的半监督谱聚类算法明显优于 K-means 方法, K-means 在高维空间对稀疏数据的处理并不好, 而半监督谱聚类算法很好地解决了这一问题, 并且谱聚类算法能够适应网络流量分布广、差异大、维度高的特点. 图 4 展示了半监督谱聚类算法对数据集S进行聚类后的分布情况, 可以看出在簇数为 9 的情况下, 仍能很好地区分出各个类别.

表 1 半监督图谱聚类实验结果

检测类型	K-means	本文
Normal	0.78	0.92
Neptune	0.66	0.86
Pod	0.58	0.94
Satan	0.63	0.76
Nmap	0.42	0.84
Warezmaster	0.36	0.88
Spy	0.45	0.79
Sqlattack	0.33	0.86
Rootkit	0.49	0.82

3.4 本文模型整体性能评估

首先实验利用自生成数据集对本文提出的模型做了整体性的评估, 实验结果如表 2 所示. 可以看出本文提出的模型总体上在对各个类别的识别上具有较高的识别率, 均在 92% 以上, 最高可达 99%, 在各个类别的

预测上表现都很出色, 并且性能均衡稳定, 初步证明了本文方法在未知恶意流量识别上的有效性.

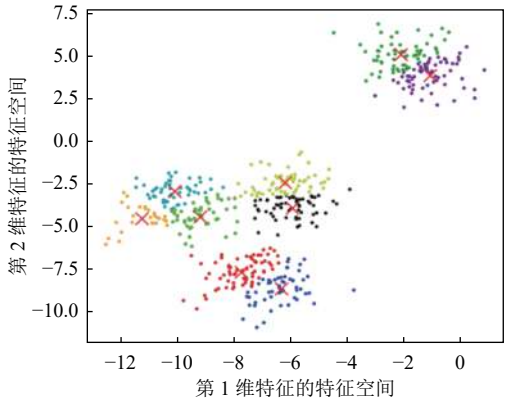

图 4 半监督谱聚类散点分布图

表 2 本文方法在自生成数据集上的整体性能评估

检测类型	ACC	Precision	Recall	FPR	*F*1-score
Normal	0.9609	0.9781	0.9805	0.1242	0.9792
Neptune	0.9731	0.987	0.9844	0.1135	0.9856
Pod	0.9648	0.9774	0.9855	0.2352	0.9814
Satan	0.9735	0.977	0.9953	0.1565	0.9860
Nmap	0.9766	0.9841	0.9914	0.1648	0.9877
Warezmaster	0.9425	0.9425	0.9845	0.1468	0.9630
Spy	0.9234	0.9065	0.9465	0.2016	0.9260
Sqlattack	0.9346	0.9168	0.9515	0.1674	0.9338
Rootkit	0.9586	0.8916	0.9457	0.1598	0.9178

基于 Multi-SMOTE 过采样的流量处理方法能够提高流量特征质量, 从而提高恶意未知流量识别的准确率. 基于半监督谱聚类的未知流量筛选方法是为了专门筛选出未知流量以避免其他已知恶意流量对结果的影响. 这两个步骤对于未知恶意流量的检测起到了非常重要的作用, 为了验证其有效性, 实验利用 3 种数据训练了集成 SVM 模型. 数据 1 为原始数据, 即为没有经过 Multi-SMOTE 过采样处理和基于半监督谱聚类的未知流量筛选的数据, 表示为 Data1. 数据 2 为仅经过基于半监督谱聚类的未知流量筛选的数据, 表示为 Data2. 数据 3 为同时经过 Multi-SMOTE 过采样处理和基于半监督谱聚类的未知流量筛选的数据, 表示为 Data3. 在实验的过程中, 同样均利用 Bagging 思路训练集成 SVM 分类器, 分类结果如表 3–表 5 所示.

从表 3–表 5 的结果可以综合看出, 在不同的数据集上, 3 种数据的总体表现趋势是一致的. 使用原始数据训练出的集成 SVM 分类器, 即使使用了 Bagging 思想, 也几乎没有分类效果. 经过基于半监督谱聚类的未知流量筛选后的数据, 在训练集成 SVM 分类器的时候有一点效果, 但是 *F*1-score 最好值也只有 0.806, 说明特征的质量对实验结果有很重要的影响. 经过本文模型设计的所有环节处理后的数据, 再利用 Bagging 思想训练出的集成 SVM 模型在综合评价上表现优异, 验证了本文提出模型的每一个环节都是不可缺失的, 环环紧扣, 最终实现了对未知恶意流量的准确识别.

表 3 集成 SVM 分类器在 3 种数据上的表现 (自生成数据集)

数据	ACC	Precision	Recall	FPR	*F*1-score
Data1	0.4545	0.5	0.45	0.5454	0.474
Data2	0.727	0.8	0.5	0.273	0.615
Data3	0.991	0.992	0.992	0.009	0.99

表 4 集成 SVM 分类器在 3 种数据上的表现 (UCI 数据集)

数据	ACC	Precision	Recall	FPR	*F*1-score
Data1	0.477	0.522	0.5	0.523	0.51
Data2	0.768	0.785	0.79	0.232	0.789
Data3	0.945	0.92	0.983	0.05	0.952

表 5 集成 SVM 分类器在 3 种数据上的表现 (KDD'99)

数据	ACC	Precision	Recall	FPR	*F*1-score
Data1	0.5	0.54	0.583	0.5	0.56
Data2	0.78	0.78	0.833	0.22	0.806
Data3	0.977	0.967	0.992	0.023	0.979

为了进一步验证本文方法的先进性, 我们选取了文献 [25,27,29,34] 中的方法进行了对比实验. 结果如表 6 所示, 我们选取 KDD'99 数据集作为实验数据. 其中, 文献 [27,29,34] 使用了 KDD'99 的数据集, 文献 [25] 使用了 UCI 数据集. KDD'99 数据集公布于 1999 年, 是网络流量检测领域中使用最广泛的数据集, 也是目前比较权威的数据集. KDD'99 数据集是由 DARPA98 数据集经过数据挖掘和预处理后得到的, 每条记录由 41 个属性组成, 分成 4 大类共 39 种攻击类型.

实验首先利用文献 [25] 提出的结合决策树和神经网络模型方法进行了对比, 由于决策树采用规则, 在一定程度上依赖数据本身结构, 缺少了灵活性, 本文提出的方法对各类恶意流量的适应性更强, 所以效果优于文献 [25] 提出的方法. 随后, 本文对同在 KDD'99 数据集进行实验的文献 [27,29,34] 进行了对比, 这 3 个工作的实验是针对分类器做的改进, 通过多种算法相结合的方式, 从而提升分类器的性能, 但忽视了数据分布的情况和数据集规模大小的问题. 本文提出的方法利用半监督形式来适应数据分布的问题. 从表 3 中可以看出, 本文方法在准确率 (ACC), 精确率 (precision), 召回率 (recall) 上均优于其他方法.

表 6 本文方法与其他方法的性能对比

指标	文献[25]	文献[34]	文献[27]	文献[29]	本文方法
ACC	0.953	0.973	0.988	0.851	0.977
Precision	0.898	0.896	0.917	0.917	0.967
Recall	0.851	0.849	0.997	0.811	0.992
FPR	—	0.152	0.013	0.097	0.023
*F*1-score	0.874	0.814	0.955	0.861	0.979

在实际的应用中, 最重要的目标是尽可能发现异常行为, 提高检测的准确率和召回率, 而 0.1 的误报率在可接受的范围内.

为了测试本文方法的泛化能力, 我们利用本文的方法在自生成数据集*S*, UCI 数据集和 KDD'99 数据集上分别做了实验, 并根据 *F*1-score 对性能进行了评估. 图 5 展示了 3 个数据集上 *F*1-score 的大小, 由于*S*是本文提出的私有数据集, 与模型兼容性更高, 所以 *F*1-score 达到了 0.99, 另外两个数据集的 *F*1-score 在 0.95 以上.

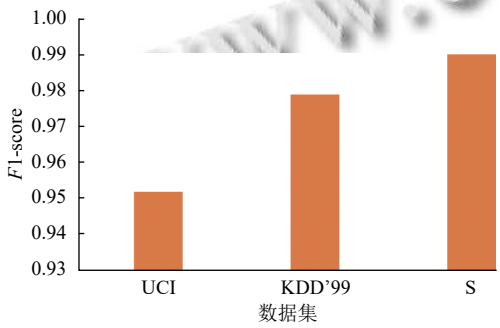

图 5 本文方法在不同数据集上的性能比较

综上所述, 从整体实验结果来看, 本文所提方法的整体优势明显, 且具有良好的泛化能力.

4 结论

本文针对真实网络环境中存在恶意流量数据极度不平衡的问题, 研究具有高综合性能的未知恶意流量检测方法. 综合考虑传统恶意流量检测方法中存在样本不均衡性、样本分布适应性不足等问题, 采用机器学习方法对未知恶意流量进行检测.

首先, 本文提出基于 Multi-SMOTE 过采样的流量处理方法, 为后续未知攻击检测步骤提供高质量的数据集, 降低数据质量导致的检测误差; 然后提出一种基于半监督谱聚类的未知流量筛选方法, 从混合流量中精准识别未知流量. 最后, 利用前面提出的过采样方法和未知流量筛选处理后的特征, 基于 Bagging 思想, 训练了能够识别未知恶意流量的集成 SVM 分类器. 实验结果表明, 本文所提出的基于集成 SVM 与 Bagging 的未知恶意流量检测模型在综合评价 (*F*1-score) 方面优于目前同类未知恶意流量检测方法. 并且, 本文所提方法在不同数据集上的性能评估结果显示, 所提方法具有良好的泛化能力.

参考文献

1 CNCERT. 2021 年上半年我国互联网网络安全监测数据分析报告. 北京: 国家互联网应急中心, 2021.
2 Bauer JM, Dutton WH. The new cybersecurity agenda: Economic and social challenges to a secure internet. Washington: World Bank, 2015.
3 Tumer D, Entwisle S. Symantec internet security threat report trends (Volume XI). Technical Report, Cupertino: Symantec Inc., 2006.
4 深信服千里目安全实验室. 2019 年上半年网络安全态势报告. https://www.freebuf.com/articles/paper/210122.html. (2019-08-06).
5 Aldweesh A, Derhab A, Emam AZ. Deep learning approaches for anomaly-based intrusion detection systems: A survey, taxonomy, and open issues. Knowledge-based Systems, 2020, 189: 105124. [doi: 10.1016/j.knosys.2019.105124]
6 Chawla NV, Bowyer KW, Hall LO, *et al*. SMOTE: Synthetic minority over-sampling technique. Journal of Artificial Intelligence Research, 2002, 16: 321–357. [doi: 10.1613/jair.953]
7 Han H, Wang WY, Mao BH. Borderline-SMOTE: A new over-sampling method in imbalanced data sets learning. Proceedings of the 2005 International Conference on Advances in Intelligent Computing. Hefei: Springer, 2005. 878–887.
8 He HB, Bai Y, Garcia EA, *et al*. ADASYN: Adaptive synthetic sampling approach for imbalanced learning. 2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence). Hong Kong: IEEE, 2008. 1322–1328.
9 Zhang X, Ma D, Gan L, *et al*. CGMOS: Certainty guided minority oversampling. Proceedings of the 25th ACM International on Conference on Information and Knowledge Management. Indianapolis: ACM, 2016. 1623–1631.
10 Tang B, He HB. KernelADASYN: Kernel based adaptive synthetic data generation for imbalanced learning. 2015 IEEE Congress on Evolutionary Computation (CEC). Sendai: IEEE, 2015. 664–671.
11 Zhu TF, Lin YP, Liu YH. Synthetic minority oversampling technique for multiclass imbalance problems. Pattern Recognition, 2017, 72: 327–340. [doi: 10.1016/j.patcog.2017.

07.024]
12 黄海松, 魏建安, 康佩栋. 基于不平衡数据样本特性的新型过采样 SVM 分类算法. 控制与决策, 2018, 33(9): 1549–1558. [doi: 10.13195/j.kzyjc.2017.0649]
13 刘根城. 基于模糊规则的过采样技术研究 [硕士学位论文]. 西安: 西安电子科技大学, 2019. [doi: 10.27389/d.cnki.gxadu.2019.001265]
14 Liu SG, Zhang J, Xiang Y, *et al*. Fuzzy-based information decomposition for incomplete and imbalanced data learning. IEEE Transactions on Fuzzy Systems, 2017, 25(6): 1476–1490. [doi: 10.1109/TFUZZ.2017.2754998]
15 Liang XW, Jiang AP, Li T, *et al*. LR-SMOTE—An improved unbalanced data set oversampling based on K-means and SVM. Knowledge-based Systems, 2020, 196: 105845. [doi: 10.1016/j.knosys.2020.105845]
16 Konstantinou E. Metamorphic Virus: Analysis and Detection. London: University of London, 2008: 93.
17 李剑. 恶意软件行为分析及变种检测技术研究 [硕士学位论文]. 杭州: 杭州电子科技大学, 2009.
18 韩晓光, 曲武, 姚宣霞, 等. 基于纹理指纹的恶意代码变种检测方法研究. 通信学报, 2014, 35(8): 125–136. [doi: 10.3969/j.issn.1000-436x.2014.08.016]
19 Al-Yaseen WL, Othman ZA, Nazri MZA. Multi-level hybrid support vector machine and extreme learning machine based on modified K-means for intrusion detection system. Expert Systems with Applications, 2017, 67: 296–303. [doi: 10.1016/j.eswa.2016.09.041]
20 Bailey M, Collins C, Sinda M, *et al*. Intrusion detection using clustering of network traffic flows. 2017 18th IEEE/ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD). Kanazawa: IEEE, 2017. 615–620.
21 Vokorokos L, Baláž A, Trelová J. Distributed intrusion detection system using self organizing map. IEEE 16th International Conference on Intelligent Engineering Systems. Lisbon: IEEE, 2012. 131–134.
22 Hassan MM, Gumaei A, Alsanad A. A hybrid deep learning model for efficient intrusion detection in big data environment. Information Sciences, 2020, 513: 386–396. [doi: 10.1016/j.ins.2019.10.069]
23 Benisha RB, Ratna SR. Detection of data integrity attacks by constructing an effective intrusion detection system. Journal of Ambient Intelligence and Humanized Computing, 2020, 11(11): 5233–5244. [doi: 10.1007/s12652-020-01850-1]
24 Feng YX, Kang YY, Zhang H, *et al*. FAFS: A fuzzy association feature selection method for network malicious traffic detection. KSII Transactions on Internet and Information Systems, 2020, 14(1): 240–259.
25 Bondarenko A, Aleksejeva L, Jumutc V, *et al*. Classification tree extraction from trained artificial neural networks. Procedia Computer Science, 2017, 104: 556–563. [doi: 10.1016/j.procs.2017.01.172]
26 Li JY, Ma SS, Le T, *et al*. Causal decision trees. IEEE Transactions on Knowledge and Data Engineering, 2017, 29(2): 257–271. [doi: 10.1109/TKDE.2016.2619350]
27 Mohamed MR, Nasr AA, Tarrad IF, *et al*. Exploiting incremental classifiers for the training of an adaptive intrusion detection model. International Journal of Network Security, 2019, 21(2): 275–289.
28 Kamarudin MH, Maple C, Watson T, *et al*. A logitboost-based algorithm for detecting known and unknown web attacks. IEEE Access, 2017, 5: 26190–26200. [doi: 10.1109/ACCESS.2017.2766844]
29 Kim KM, Hong JN, Kim K, *et al*. Evaluation of ACA-based intrusion detection systems for unknown-attacks. 2016 Symposium on Cryptography and Information Security. Kumamoto: The Institute of Electronics, Information and Communication Engineers (IEICE), 2016. 1–8.
30 许勐璠, 李兴华, 刘海, 等. 基于半监督学习和信息增益率的入侵检测方案. 计算机研究与发展, 2017, 54(10): 2255–2267. [doi: 10.7544/issn1000-1239.2017.20170456]
31 Lin WC, Ke SW, Tsai CF. CANN: An intrusion detection system based on combining cluster centers and nearest neighbors. Knowledge-based Systems, 2015, 78: 13–21. [doi: 10.1016/j.knosys.2015.01.009]
32 Bamakan SMH, Wang HD, Tian YJ, *et al*. An effective intrusion detection framework based on MCLP/SVM optimized by time-varying chaos particle swarm optimization. Neurocomputing, 2016, 199: 90–102. [doi: 10.1016/j.neucom.2016.03.031]
33 Gharaee H, Hosseinvand H. A new feature selection IDS based on genetic algorithm and SVM. 2016 8th International Symposium on Telecommunications (IST). Tehran: IEEE, 2016. 139–144.
34 Aslahi-Shahri BM, Rahmani R, Chizari M, *et al*. A hybrid method consisting of GA and SVM for intrusion detection system. Neural Computing and Applications, 2016, 27(6): 1669–1676. [doi: 10.1007/s00521-015-1964-2]
35 Kumar G, Thakur K, Ayyagari MR. MLEsIDSs: Machine learning-based ensembles for intrusion detection systems—A review. The Journal of Supercomputing, 2020, 76(11): 8938–8971. [doi: 10.1007/s11227-020-03196-z]

(校对责编: 孙君艳)