广东海洋大学 Guangdong Ocean University

2022 级物联网工程专业课程论文

《信息资源检索与应用》课程论文

网络入侵检测中类别平衡与特征选择方法的应用

Course Paper for Information Resources Retrieval and Application

专业：物联网工程专业

班级：物联 1224

学号：202211672411

姓名：黎川滔

考核项目	支撑课程目标	分值	考核标准点	得分
论文选题及难度	课程目标 3：具备一定的国际视野，对物联网行业的国际发展趋势有一定的了解，能够就复杂物联网工程问题与业界同行及社会公众进行沟通和交流	11	（1）课程论文题目是物联网专业相关的最新技术； （2）论文题目要有一定的难度和深度。 （3）论文参考文献是否有 SCI 和 EI 检索的国际期刊会议论文。 （4）撰写论文要积极主动和老师和同学沟通和交流	
论文文献收集和整理	课程目标 2：论文能够运用图书馆、互联网、数据库等资源进行信息检索、资料查询和归纳总结	39	（1）文献的检索和收集是否齐全和准确 （2）文献的分析和加工是否到位	
论文的撰写	课程目标 1：论文能够通过文献研究，深入分析复杂物联网工程问题，以获得有效的结论	29	（1）论文内容全面性、逻辑清晰性、章节划分合理性。 （2）论文分析是否详实具体，论文归纳总结是否到位。 （3）论文是否得到有效结论。	
论文格式及参考文献	课程目标 4：具有自主学习和终身学习的意识，能够阅读和理解物联网专业文献，主动学习专业知识和技术	21	（1）论文格式要规范，包括字体，图表，公式等 （2）参考文献引用及格式要规范。 （3）论文写作和提交的积极性和主动性。	
综合得分				

网络入侵检测中类别平衡与特征选择方法的应用

黎川滔[1]，梁祖仲[1,†]

(1. 广东海洋大学 数学与计算机学院，广东湛江 524000)

摘　要：随着互联网服务的广泛普及，尤其是在应对新型攻击和破坏方面，服务提供商和用户面临着日益严峻的系统安全保护挑战。网络入侵检测系统通过分析网络数据包来识别潜在威胁。然而，在处理特征冗余和类别不平衡严重的数据集时，提升入侵检测的预测精度已成为亟待解决的关键问题。本研究对网络入侵检测系统中的类别平衡和特征选择技术进行了全面系统的分析。本研究回顾和总结了近来提出的相关文献，通过比较不同研究中的各种方法，分析了其优点、局限性和潜在的优化策略。此外，本文还指出了当前网络入侵检测领域面临的主要挑战和问题，并对未来的研究方向提出了见解，旨在通过提供可操作的建议来指导和激励这一领域的未来发展。

关键词：网络入侵检测，统计机器学习，不平衡学习，特征选择、

中图分类号：TP393

Application of Class Balance and Feature Selection in Network Intrusion Detection

Li Chuantao[1], Liang Zuzhong[1,†]

(1. College of Mathematics and Computer, Guangdong Ocean University, Zhanjiang Guangdong, 524000, China)

Abstract: As internet services become increasingly widespread, service providers and users face escalating challenges in protecting system security, particularly against emerging attacks and disruptions. Network Intrusion Detection Systems play a critical role in identifying potential threats by analyzing network packets. However, improving the predictive accuracy of intrusion detection systems has become a pressing issue, especially when dealing with datasets characterized by feature redundancy and severe class imbalance. This study offers a comprehensive and systematic analysis of category balance and feature selection techniques in network intrusion detection systems. It critically reviews and synthesizes recent literature, highlighting the advantages, limitations, and potential optimization strategies by comparing various approaches across different studies. Furthermore, the paper identifies key challenges and issues currently faced in the field of network intrusion detection and provides insights into future research directions. The study aims to guide and inspire future advancements in this realms by offering actionable recommendations.

Key words: Network Intrusion Detection, Statistical Machine Learning, Imbalanced Learning, Feature Selection

0　引言

现如今，互联网在社会和经济生活中的作用越来越重要，互联网新兴技术的不断涌现为产业革新和经济发展赋予了全新动能。然而，频发的的网络攻击与非法访问为互联网在日常生活中的应用带来了不可忽视的安全隐患，使服务提供商和用户面临着日益严峻的系统安全保护挑战[1]。

2010 年，伊朗布什尔核电站设备遭受震网（Stuxnet）蠕虫攻击，高度复杂的恶意代码及多个 0day 漏洞造成约 1000 台铀浓缩离心机故障[2]。2015 年 12 月 23 日，乌克兰伊万诺弗兰科夫斯克地区遭受高级可持续性威胁（APT）攻击导致大规模停电，该攻击是由于电力系统感染携带 Killdisk 组件的恶意木马 BlackEnergy 所导致[3]。2020 年 12 月 13 日，FireEye 公司发布报告，声称 SolarWinds 旗下的软件遭到供应链攻击，黑客通过篡改源代码并添加后门实施网络攻击[4]。2022 年 9 月，根据 360 公司最新溯源研究披露，美国国家安全局（NSA）针对某大学发起上千次网络攻击活动，通过精心构造的武器库控制内网设备并窃取高价值数据，再利用 49 台跳板机和多个 0day 漏洞（譬如 Extremeparr 和 Ebbisland）开展定向攻击，造成巨大的损失[5]。这些攻击事件表明，网络入侵给全球政治、经济、科技、教育、文化各方面带来严重影响，我国作为主要受害国正面临严峻的网络攻击威胁[6]。因此，建立一个高效稳健的网络侵检测系统（NIDS）以保护联网设备免受网络攻击，则显得尤为重要。网络入侵检测系统旨在通过监视网络流量，在检

基金项目：广东海洋大学教育创新计划资助项目（202410566025）。

作者简介：黎川滔（2004-），男，广东汕尾人，本科生，主要研究方向为数据挖掘，统计机器学习；梁祖仲（1994-），男，广东茂名人，助理讲师，硕士研究生，主要研究方向为图像增强和去模糊，1449165991@qq.com。

测到的任何恶意或可疑活动时发出警报。

网络入侵检测系统主要分为两大类，一种是基于异常的网络入侵检测系统，另一种是基于标签的网络入侵检测系统。其中，基于标签的网络入侵检测系统由于其通过匹配已知攻击模式来精准识别恶意行为而具有较高的灵活性，从而在实际应用中更为常见。诚然，现有的基于标签的网络入侵检测系统面临两个主要挑战：首先，这些模型的训练数据集通常存在攻击类别高度不平衡的问题，导致模型在样本数量较少的攻击类别上的检测性能不足。其次，这些数据集的特征冗杂繁多，不仅会降低机器学习模型的计算速度，还会导致模型在学习过程中过拟合，进而在真实网络环境中出现较高的误报率。

针对攻击类别不平衡的问题，大多数文献采用了统计方法或对抗生成方法。统计方法主要分为过采样和欠采样。欠采样直接从多数类中随机移除一部分样本，使得两类样本数量相等。当多数类样本非常丰富时，这种方法可能导致模型无法捕捉到足够的多数类特征。过采样是通过增加少数类样本的数量来平衡类别分布，这种方法保留了原始数据中的信息，因而在网络入侵检测系统更为适用。由于统计方法生成样本的策略为特征空间中寻找距离最短的样本，这将导致合成样本和真实样本之间难以确定清晰的决策边界，从而使得合成样本的质量易受噪声的影响。然而，生成对抗方法如 GAN（生成对抗网络）通过训练生成器神经网络和判别器神经网络进行博弈，从而在互相反馈中不断提升自身性能。最终，生成对抗方法生成的样本更符合真实样本分布趋势。

针对特征冗余问题，繁杂不相关的特征不仅扩大了数据集规模，还增加了模型过拟合的风险，导致在真实网络入侵检测任务中存在较高的假阳性率。现有的网络入侵检测模型大多采用以下三种特征选择方法：过滤式方法、嵌入式方法和包装式方法。过滤式方法计算效率高，通过如相关性、方差或信息增益等评估指标快速选择与心脏病风险最相关的特征。诚然，许多过滤式方法都是基于单个特征的统计特性而建立的，并没有考虑特征之间的相互关系。与其他特征选择方法不同的是，如 LASSO 和弹性网络等嵌入式方法，将特征选择过程建立在模型的学习上。尽管嵌入式方法在线性空间上表现出较为优越的性能，但其处理非线性特征组合的能力有限，不能充分捕捉复杂的数据模式。包装式方法使用的预警模型遍历特征子集，通过设计的损失函数度量每个特征子集的质量，因此可以与任何统计机器学习方法结合。然而，遍历每一种特征组合属于 NP-Hard 问题，这使得包装式方法需要消耗巨大的计算资源。

为此，本研究面向网络入侵检测中类别平衡与特征选择方法开展系统、深入、全面地归纳和总结，通过 Google Scholar、ScienceDirect、IEEE Xplore、Springer Link 等数据库检索相关文献进行整理，并展望未来研究方向。本研究的主要贡献如下：

(1) 本研究分别对网络入侵检测系统常用的类别平衡方法和特征选择方法进行全面地总结、关联、归纳和分析。在方法介绍方面，本研究给出具体的定义和介绍，并详细总结各种方法的优缺点。

(2) 本研究系统地回顾网络入侵检测系统关于类别平衡方法和特征选择方法的相关研究工作，全面概述不同文献对于检测系统的实现过程以及设计特点，探索和对比相关工作的优缺点。

(3) 本研究针对现有的研究，总结并展望网络入侵检测系统的研究趋势及存在挑战，给出未来研究方向的建议。

文章剩余部分的结构为：第一章介绍了网络入侵检测领域常见的几种类别平衡方法，简要概述了算法流程和方法优缺点；第二章介绍了网络入侵检测领域常见的三类特征选择方法，并指出它们的优势和缺陷；第三章全面总结分析了网络入侵检测类别平衡方法的应用综述；第四章全面总结分析了网络入侵检测特征选择方法的应用综述；第五章讨论当前的问题及展望未来研究方向；最后，本研究在第六章进行总结。

1 类别平衡方法概述

1.1 Tomek-Links

Tomek-Links 是一种处理不平衡数据集的欠采样技术，其核心思想是识别并移除数据集中的 Tomek 链接，这些链接由属于不同类别但彼此之间距离非常近的样本对构成。如果样本x属于一个类别，而样本y属于另一个类别，且x和y互为最近邻，并且它们之间的距离$d(x,y)$满足一定条件，那么这两个样本就构成了一个 Tomek 链接[7]。

具体而言，Tomek-Links 首先会计算数据集中所有样本之间的距离，然后找到那些互为最近邻且属于不同类别的样本对。这些样本对被称为 Tomek 链接。接着，从数据集中移除这些 Tomek 链接中的样本，以减少多数类样本的数量，从而在一定程度上平衡数据集的类别分布。这种方法有助于区分少数类和多数类，并且通过移除那些可能对分类器决策边界产生负面影响的样本，可以提高分类器的性能

Tomek-Links 方法的优点在于它能够识别并移除那些可能导致分类器误分类的样本，从而提高模型的泛化能力。然而，Tomek-Links 可能因为其移除样本而导致信息的丢失。此外，Tomek-Links 方法主要适用于二分类问题。

1.2 SMOTE

SMOTE（Synthetic Minority Over-sampling Technique，合成少数类过采样技术）是一种过采样技术，用于解决类别不平衡问题。在该方法中，通过生成合成数据来增加少数类样本的数量，从而实现过采样。SMOTE 合成样本是少数类的两个相似样本(s, s^R)的线性组合，并被如下式子定义[8]：

$$n = s + d \cdot (s^R - s), \quad 0 \le d \le 1, \tag{1}$$

上式中，n表示合成样本，s表示少数类中的一个原始样本，s^R表示从s最近邻中随机选择的一个样本，d用于控制合成样本在s和s^R之间的线性插值程度。

具体而言，SMOTE 首先从少数类中随机选择一个样本，然后在特征空间中找到该样本的两个最近邻样本。接着随机选

择其中一个近邻样本，并计算该样本与近邻样本之间特征值的差值，将这个差值乘以一个介于 0 和 1 之间的随机数，并将结果加到原始样本的特征值上。通过这种方式，在原始样本和其邻居之间的决策边界上生成合成样本。根据所需的过采样数量，SMOTE 还会从k近邻中随机选择多个近邻样本来生成合成样本，其中k的值是一个超参数。

SMOTE 由于其合成样本点特点，相对于随机采样能够提高模型对少数类的识别能力。然而，SMOTE 生成的合成样本可能过于集中在少数类样本的密集区域，且容易受到噪声的影响，不适合应用于高维度数据集。

1.3 ADASYN

ADASYN（Adaptive Synthetic Sampling）是 SMOTE 的改进版本，它通过判断少数类样本的学习难度来确定样本的加权分布，从而在难以学习的少数类样本周围生成更多的合成样本[9]。ADASYN 按如下公式计算数据集的不平衡度d：

$$d = \frac{m_s}{m_l}, \tag{2}$$

上式中，m_s表示少数类样本的数量，m_l表示多数类样本点数量。

如果d小于最大容忍不平衡阈值d_{th}，则按如下公式计算需要生成的样本数量G：

$$G = (m_l - m_s) \times \beta, \tag{3}$$

上式中，β表示所需平衡度，用于设置生成样本后合成数据集的期望平衡度。

对每个少数类样本x_i，计算其k近邻中属于多数类的样本数量r_i，并按如下公式计算每个少数类样本点权重$\hat{r}_i$：

$$\hat{r}_i = \frac{r_i}{\sum_{i=1}^{m_s} r_i}. \tag{4}$$

根据权重$\hat{r}_i$和需要生成的样本数量G，为每个少数类样本生成相应数量的新样本。其中，新样本通过在少数类样本x_i与其k近邻中的少数类样本之间进行插值生成。

具体而言，ADASYN 首先计算数据集的不平衡度，若不平衡度小于最大容忍不平衡阈值，则计算需要生成的样本数量。对于每个少数类样本，计算其k近邻中属于多数类的样本数量和每个少数类样本点权重。接着根据权重和需要生成的样本数量，为每个少数类样本生成相应数量的合成样本。

ADASYN 算法通过自适应地为难以学习的少数类样本生成更多的新样本，有效地解决了不平衡数据集中的类别不平衡问题，提高了模型在少数类上的预测性能。然而，这种基于k近邻方法产生的样本会扩宽少数类样本集合的边界，使得产生的新样本容易与多数类样本混淆。

1.4 GAN

利用二进制交叉熵作为其损失函数的 GAN 模型通常被称为 Vanilla GAN 或简称为 GAN。在 GAN 模型中，判别器的任务是对合成数据和真实数据进行辨别，其输出概率$D(x)$表示合成样本x来自真实数据分布而非生成器生成的概率。通常，p判别器的损失函数使用如下式子定义[10]：

$$L_D = -\mathbb{E}_{x\sim p_{data}(x)}[\log D(x)] - \mathbb{E}_{x\sim P_{gen}(x)}\big[\log\big(1 - D(x)\big)\big], \tag{5}$$

上式中，$-\mathbb{E}_{x\sim p_{data}(x)}[\log D(x)]$表示判别器关于真实数据样本的对数似然的负期望值，这是判别器针对真实样本的损失部分，模型将最小化该部分。$\mathbb{E}_{x\sim P_{gen}(x)}[\log\big(1 - D(x)\big)]$表示对于合成样本$x$，判别器未能正确识别这些样本为“假”的期望损失，模型将最大化这一部分，使得判别器难以区分合成数据和真实数据。

1.5 Wasserstein GAN (WGAN)

在标准的 GAN 中，训练过程试图最小化生成器和判别器网络的二进制交叉熵损失，从而达到的目的。然而，标准 GAN 的训练可能具有挑战性，因为它们会遇到模式崩溃和梯度消失等问题。Wasserstein GAN 通过使用 Wasserstein 距离作为损失函数来解决其中的一些挑战。该距离量化了将一个概率分布转换为另一个概率分布所需的最少消耗[11]。

Wasserstein 距离被分布P和分布Q所定义：

$$\text{Wasserstein}(P, Q) = \min_{f\in\prod(P,Q)} \sum_{i\in P} \sum_{j\in Q} d(i,j)\cdot f(i,j), \tag{6}$$

上式中，$\text{Wasserstein}(P, Q)$表示分布P和分布Q的差异。$d(i,j)$表示从分布P中的数据点i到分布Q中的数据点j的距离，其衡量了两个数据点之间的相似性或差异性。$f(i,j)$表示从分布P中的数据点i到分布Q中的数据点j需要的移动质量，其描述了如何将分布P中的质量重新分配到分布Q中，以使两个分布尽可能接近。

WGAN 利用 Wasserstein 距离作为 GAN 中的损失函数，在训练过程中引入了连续和平滑的梯度，从而增强了训练稳定性并减轻了梯度消失等问题。此外，Wasserstein 损失为训练评估提供了更多信息的度量。通过量化合成数据分布与真实数据分布的紧密程度，为生成器提供了更清晰的目标：合成数据不仅看起来真实，而且能更准确地捕获层数据数据的分布。

1.6 Wasserstein GAN with Gradient Penalty(WGAN-GP)

在 WGAN 中，权重裁剪用于确保判别器的输出满足 Lipschitz 连续性约束，即判别器的输出变化不会超过输入变化的一定倍数。然而，由于裁剪操作会破坏梯度的传播，WGAN 容易出现模型难以收敛的问题。针对这个问题，WGAN-GP 通过引入梯度惩罚来替代权重裁剪。

具体而言，WGAN-GP 在损失函数中加入了一个梯度惩罚项，该惩罚项确保判别器的输出变化与输入变化之间的比率接近 1，梯度惩罚项定义如下[12]：

$$\text{GP} = \lambda\mathbb{E}_{\hat{x}\sim P_{\hat{x}}}[(\|\nabla_{\hat{x}} D(\hat{x})\|_2 - 1)^2], \tag{7}$$

上式中，$\hat{x}$表示从真实数据分布P_{data}和合成数据分布P_{gen}之间的线性插值点，$D(\hat{x})$表示判别器的输出，$\nabla_{\hat{x}} D(\hat{x})$表示判别器输出关于$\hat{x}$的梯度，$\lambda$是一个超参数，用于控制梯度惩罚的强度。针对上述的几种统计方法和生成方法，本研究对其优缺点总结如表 1 所示。

表 1 统计方法和生成方法在类别平衡的性能对比

Table 1 Performance comparison of statistical methods and generation methods in class balancing

算法名称	优点	缺点
Tomek-Links	通过去除决策边界上的样本来减少类间重叠 用于清理数据中的噪声和边界不明确的样本	只对少数类和多数类边界上的样本进行处理 仅适用于二分类，且不适用于高维数据
SMOTE	实现简单 基于最近邻思想生成样本	基于线性插值易引入噪声 不适用于高维数据
ADASYN	能够根据数据的分布密度生成样本 有效地提高模型对难分类样本的敏感性	容易扩宽多数类与少数类的决策边界 容易会生成冗余样本，导致过拟合
GAN	不仅适用于少数类样本生成，还可以用于数据增强任务 强大的生成能力，能够生成高质量的样本	容易出现崩溃现象 需要较大的计算资源，且参数调优难度较大
WGAN	基于 Wasserstein 距离，减少了模式崩溃的风险 能够生成更加平滑和逼真的数据分布	在高维数据上仍可能遇到收敛问题 训练过程较慢，参数选择敏感
WGAN-GP	引入梯度惩罚进一步提高了训练的稳定性，避免 WGAN 中对判别器过度限制的缺陷 能够生成更加逼真且多样化的样本	梯度惩罚系数等参数的调节要求较高 需要较长的训练时间

2 特征选择方法概述

2.1 过滤法

过滤式方法计算效率高，通过如方差、卡方统计量或互信息等评估指标快速选择与攻击类别最相关的特征[13]。诚然，过滤式特征选择方法在网络入侵检测领域暴露出了许多缺点。一方面，许多过滤式方法都是基于单个特征的统计特性而建立的，并没有考虑特征之间的相互关系；另一方面，过滤法通常假设特征与目标变量之间的关系是线性的或简单可加的，故这类方法无法准确捕捉其网络流量特征之间的重要性[14]。

2.2 方差分析

方差分析（ANOVA）适用于分类问题中的特征选择，对于每个特征，方差分析计算该特征对不同类别样本均值差异的影响，即组间方差与组内方差的比例（F 值）。较大的 F 值意味着该特征能够更好地区分不同的类别，使用如下公式计算：

$$F=\frac{\frac{\sum_{j=1}^{J}N_j\left(\overline{x}_j-\overline{x}\right)^2}{(J-1)}}{\frac{\sum_{j=1}^{J}\left(N_j-1\right)s_j^2}{(N-1)}}, \tag{8}$$

上式中，N_j表示第j个特征的样本数量，$\bar{x}_j$表示第j个特征的平均值，$\bar{x}$表示所有特征的总体平均值，s_j^2表示第j个特征的样本方差，N表示所有特征的样本总数。

2.2.1 卡方检验

卡方检验（Chi-Squared Test）适用于离散型特征与类别标签之间的关联性分析。卡方检验通过比较观测频数与期望频数之间的差异来判断某个特征是否与类别标签有关联。其中，卡方统计量使用如下公式计算：

$$\mathcal{X}^2=\sum\frac{\left(o_j-e_j\right)^2}{e_j}, \tag{9}$$

上式中，o_j表示第j个特征的观测频数，e_j表示第j个特征的期望频数。

2.2.2 互信息

互信息（Mutual Information）度量了两个随机变量之间的依赖程度，对线性和非线性依赖均可度量。在特征选择过程中，互信息能够找到那些与目标变量之间存在非线性关系的特征。其中，互信息使用如下公式计算：

$$I(X;Y)=H(Y)-H\left(\frac{Y}{X}\right), \tag{10}$$

上式中，$I(X;Y)$表示变量X和变量Y之间的互信息，$H(Y)$表示变量Y的熵，$H\left(\frac{Y}{X}\right)$表示在给定变量$X$的条件下变量$Y$的条件熵，其计算方式分别如公式(11)和公式(12)所示。

$$H(Y)=-\sum_{y\in Y}p\left(y\right)\log p(y), \tag{11}$$

$$H\left(\frac{Y}{X}\right)=-\sum_{x\in X}\sum_{y\in Y}p\left(x,y\right)\log p(y|x). \tag{12}$$

2.3 嵌入法

嵌入法将特征选择过程建立在模型的学习上，使用模型参数来确定哪些特征对目标变量最有用[15]，主要分为基于惩罚项的模型和基于树的模型。

基于惩罚项的模型通过对线性模型引入正则化从而抑制过拟合，如 L1 正则化具有稀疏解的特性，故其天然具备特征选择的能力。基于树的模型在树的构建过程中计算如不纯度下降量等指标评估特征对目标变量的影响程度，从而选择出与目标变量最相关的特征。

然而，尽管嵌入式方法在线性空间上表现出较为优越的性能，但其处理非线性特征组合的能力有限，不能充分捕捉复杂的数据模式。除此之外，由于嵌入式方法紧密耦合于特定模型，故从一个模型转移到另一个模型时，特征选择结果难以保持普适性[16]。

2.3.1 LASSO 回归

LASSO 回归是适用于处理高维复杂共线性数据，其基本思想是在拟合广义线性回归的基础上加上了 L1 范数惩罚项。相比于岭回归的 L2 范数惩罚项，LASSO 回归可以更好地将不

重要的变量系数压缩为 0，因此被广泛应用于特征降维和抗过拟合上。其中，构建的惩罚函数如下：

$$\min_{\beta}\left\{\frac{1}{2n}\|y-X\beta\|_2^2+\lambda\|\beta\|_1\right\}, \tag{13}$$

上式中，$\frac{1}{2n}\|y-X\beta\|_2^2$是标准的线性回归损失函数，衡量预测值与实际观测值之间的均方误差。$\lambda\|\beta\|_1$表示 L1 范数正则化，λ是正则化参数，控制惩罚力度，$\|\beta\|_1$表示所有系数绝对值之和。

2.3.2 随机森林

随机森林（Random Forest）是一种通过构建多棵决策树来提高分类精度并降低过拟合风险的集成学习方法。随机森林的核心思想是使用一组决策树对输入数据进行训练和预测，然后将所有树的结果进行聚合以做出最终的预测每棵决策树T_i都是在从原始数据集中抽取的 bootstrap 样本D_i上进行训练的，并且在每个节点上，都会随机选择一个特征子集来划分数据，从而在树之间引入可变性。

对给定输入x的预测是通过聚合森林中所有决策树的预测得到的。对于分类任务，最终的预测结果由多数投票决定，可以用如下公式表示：

$$y=\text{mode}\big(T_1(x),T_2(x),\ldots,T_m(x)\big), \tag{14}$$

上式中，$T_i(x)$表示对输入x的第i棵决策树的预测，m是森林中树的总数。在多数投票中，预测的类别y是所有树中获得最多投票的类别。

随机森林在构建决策树的过程中，每个特征在每次分裂中都会贡献一定的信息增益或基尼不纯度减少量。通过累积每个特征的贡献量，可以衡量该特征对整个模型的重要性。

2.4 包装法

包装式方法使用的预警模型遍历特征子集，通过设计的损失函数度量每个特征子集的质量，因此可以与任何统计机器学习方法结合[17]。尽管遍历每一种特征组合需要消耗巨大的计算资源，但在网络入侵检测领域中，网络流量特征之间呈现非线性、多交互的特点，相比过滤法和嵌入法，包装法可以通过损失函数的设计和搜索策略的改进，从而找到最佳的特征组合，故仍然是具有可观前景的一类特征选择方法。

包装法主要使用的是启发式算法，其中最为常见的是遗传算法、粒子群算法和蚁群算法及其改进和拓展。本研究针对四种经典启发式算法性能总结对比如表 2 所示。

2.4.1 模拟退火算法

模拟退火算法（Simulated Annealing, SA）借鉴了金属退火过程中的物理原理，通过逐渐降低“温度”来使系统达到最低能量状态，从而找到全局最优解。在特征选择中，模拟退火算法通过随机选择特征子集，计算该子集的性能，并根据 Metropolis 准则决定是否接受该子集，从而逐步优化特征选择。

尽管模拟退火算法实现简单，但其传统退火方法不具备自适应功能，容易陷入局部最优解且浪费计算资源。除此之外，模拟退火算法适用于选择离散型特征，对连续型特征的选择表现不佳。

2.4.2 遗传算法

遗传算法（Genetic Algorithm, GA）是模拟自然选择和遗传学原理的优化算法。它通过模拟基因的变异、交叉和选择过程，在特征选择中逐步逼近最优特征子集。每特征子集）在 GA 中用染色体表示，经过选择、交叉、变异等操作，形成新的种群，最终找到最优解。

遗传算法适用于高维和非线性问题，能够探索较大的特征空间，其具有较好的全局搜索能力，能够有效避免陷入局部最优解。然而，遗传算法计算复杂度较高，且对交叉率、变异率等参数的设置敏感。

2.4.3 粒子群算法

粒子群优化算法（Particle Swarm Optimization, PSO）受鸟群觅食行为启发，通过模拟粒子在搜索空间中的移动来寻找最优解。每个粒子表示一个特征子集，通过与其他粒子交流信息，逐步向全局最优解逼近。PSO 的核心为每个粒子根据自身经验和群体经验来更新速度和位置，更新公式分别如公式(15)和公式(16)所示：

$$v_i^{t+1}=wv_i^t=c_1r_1(pbest_i-x_i)+c_2r_2(gbest-x_i), \tag{15}$$

$$x_i^{t+1}=x_i^t+v_i^{t+1}, \tag{16}$$

上式中，v_i表示粒子的速度，x_i表示粒子的位置，即特征子集，$pbest_i$和$gbest$分别表示粒子的局部最优解和全局最优解，r_1和r_2表示随机数，w表示惯性权重，c_1和c_2分别表示个体加速因子和社会加速因子。

PSO 参数较少，易于实现和调整，但相对于遗传算法，其适用于连续型特征的选择，在离散型特征的选择上表现不佳。

2.4.4 蚁群算法

蚁群优化算法（Ant Colony Optimization, ACO）模拟蚂蚁觅食时的集体行为，利用信息素来引导搜索方向。在特征选择中，每只蚂蚁代表一个特征子集，根据信息素的浓度决定蚂蚁选择特征的概率，从而优化特征选择过程。相比遗传算法和粒子群算法，蚁群算法通过引入信息素机制，使得蚁群算法能够引导搜索，具有较好的收敛性。

本研究对上述三类特征选择方法：过滤法、嵌入法和包装法性能的对比总结如表 3 所示。

表 2 四种经典启发式算法性能对比

Table 2 Performance comparison of four classical heuristic algorithms

算法名称	优点	缺点
模拟退火	实现简单，参数较少 对初始解的依赖性小	高维问题表现不佳 只适用于离散特征
遗传算法	适用于大规模全局搜索 能处理连续特征和离散特征	参数难以调节 收敛速度慢
粒子群算法	低维问题收敛速度快 参数较少容易调节	容易陷入局部最优解 只适用于连续特征
蚁群算法	具有较强的自适应性 信息素传递避免局部最优	对参数初始化较为敏感 计算资源大收敛速度慢

表 3 三类特征选择方法性能对比

Table 3 Performance comparison of three feature screening methods

算法名称	优点	缺点
过滤法	简单高效，适用于低维数据	不能考虑特征之间的相关性
	独立于学习算法	只通过单一统计指标选择
嵌入法	计算效率高	不适用于高维数据
	比过滤法和包装法更加准确	不同学习算法间不具普适性
包装法	考虑标签与特征之间的关系	计算资源较大
	考虑特征之间的关系	搜索策略和损失函数难以设计

3 网络入侵检测类别平衡方法的应用综述

网络入侵检测领域现有的关于类别平衡的文献中，大多数采用了统计方法或对抗生成方法，以下部分将详细分析针对数据不平衡问题的研究论文。

Injadat 等人[18]提出了一种结合类别平衡和参数优化的机器学习模型。首先，使用 SMOTE 对少数攻击类实例进行过采样，并基于信息增益和相关性进行特征选择。接着，将去除了不相关特征后的数据集输入一个参数优化过的机器学习模型，以提升网络入侵检测系统的性能。同样，Talukder 等人[19]首先对数据集进行了多种预处理技术，包括标签编码、标准化、特征缩放及 SMOTE。随后，将所选特征输入人工神经网络（ANN）、卷积神经网络（CNN）、多层感知机（MLP）、K 最近邻（KNN）、决策树（DT）和随机森林（RF），分析不同机器学习模型的分类性能。然而，上述研究中使用 SMOTE 技术进行过采样，可能会产生无信息且有噪声的样本，同时，网格搜索的超参数调整增加了时间复杂度。

Liu 等人[20]提出了一种基于 ADASYN 过采样技术和 LightGBM 的网络入侵检测系统。首先，该文章通过数据预处理对原始数据进行标准化，以避免最大值或最小值对整体特征的影响。其次，该文章通过 ADASYN 过采样技术增加了少数类别样本，以解决训练数据不均衡导致少数攻击类别检测率不足的问题。最后，LightGBM 集成学习模型用于进一步降低系统的时间复杂度，同时确保检测的准确性。该文章通过在 NSL-KDD、UNSW-NB15 和 CICIDS2017 数据集的实验验证，结果表明 ADASYN 过采样后可以提高少数类样本的预测精度，从而提高整体准确率，同时在训练和检测过程中消耗的时间更少，优于其他现有方法。

类似地，Fu 等人[21]提出了一种流量异常检测模型，称为网络入侵检测深度学习模型（DLNID），该模型结合了注意力机制和双向长短期记忆网络（Bi-LSTM）。首先，该文章通过 CNN 提取数据流量的序列特征，然后通过注意力机制重新分配每个通道的权重，最后使用 Bi-LSTM 学习序列特征的网络。针对数据不平衡问题，该文章采用 ADASYN 对少数类样本进行样本扩展，最终形成一个相对对称的数据集，并使用改进的堆叠自动编码器进行数据降维，以增强信息融合。然而，上述研究中使用基于 ADASYN 的过采样技术，虽然能够提升少数类别的检测效果，但 ADASYN 基于距离生成样本的特点，容易扩宽多数类别与少数类别的决策边界，使得模型难以清晰地区分边界上的样本。

Kim 等人[22]研究了如何通过解决数据不平衡问题来提高基于主机的入侵检测系统的异常检测精度。该文章通过 SMOTE 和 GAN 两种方法对数据中的少数类别进行了过采样，并采用 SVM 和 CNN 进行分类。研究表明，基于 GAN 的过采样在生成样本的质量上优于 SMOTE，且 CNN 在分类任务中表现优于 SVM。Lee 等人[23]使用 GAN 模型生成少数类样本，随后使用随机森林分类器评估模型性能。首先，该文章根据每个类中样本数量的多少，将训练集划分为稀有类和非稀有类数据。随后，GAN 模型在稀有类别的训练数据上进行训练，以生成稀有类别的数据样本。然而，该文章并未给出应生成样本的稀有类别的条件，导致生成的样本有时仅属于一种稀有类别。此外，GAN 还存在模式崩溃、梯度消失以及难以实现纳什均衡等问题。

Xiong 等人[24]引入了 a-model（攻击模型）、d-model（防御模型）和 t-module（黑盒训练模块）来训练网络入侵检测系统，这不仅提高了已知攻击的检测准确性，也增强了对未知攻击的识别能力。该文章将数据集中正态样本的分布作为 a-model 和 d-model 需要学习的分布，a-model 的目标是生成欺骗 d-model 的样本，而 d-model 的目标是模型是判断输入样本是否是真实样本，所以 a-model 和 d-model 之间存在对抗关系。t-module 使用 a-model 和 d-model 对抗产生的样本训练不同类型的网络入侵检测系统，称这种网络入侵检测系统为 Adversarial Training Intrusion Detection System (ATIDS)。Kumar 等人[25]通过自动编码器对原始数据集进行降维，随后使用 Wasserstein GAN-梯度惩罚（WGAN-GP）生成高质量的合成数据样本。然而，由于 WGAN-GP 在降维后的数据集上进行训练，因此未能学习到原始数据的真实分布，只是学习了低维简化的数据分布。这在一定程度上降低了网络入侵检测系统在数据集上的性能。

Srivastava 等人[26]提出了一种基于 Wasserstein 条件生成对抗网络-梯度惩罚（WCGAN-GP）和遗传算法（GA）的深度生成模型，用于生成高质量的少数类别样本，成功缓解了少数类别的不平衡问题。该文章所提出的模型生成遵循实际数据的基础数据分布的合成样本，使用一种新颖的适应度函数来收敛遗传算法，并为分类问题生成最佳特征向量。该文章应用 NSL-KDD 和 UNSW-NB15 数据集结合生成的样本和所提出模型的简化特征集进行了广泛的实验研究，以评估所提模型的优越性。这种方法有效克服了传统条件生成对抗网络在生成少数类数据时的不足，显著提高了生成数据的质量。

4 网络入侵检测特征选择方法的应用综述

特征选择是识别最相关特征以构建稳健模型的重要环节。特征选择技术通常分为三类：过滤法、嵌入法和包装法。

Akhone 等人[27]设计了一种基于过滤法的融合入侵检测系统，通过修改的决策树（MDT）和卡方特征选择方案来区分 SCADA 系统中的良性和恶意网络事件。该文章首先对原始数据进行 Z-score 归一化处理，并使用卡方分析来识别具有最高统计相关性的特征子集，随后 MDT 对选定的特征进行最终分类。但由于无法挖掘特征的组合重要性，该方法的入侵检测性能欠佳。

Alazzam 等人[28]使用鸽子优化算法，结合适应离散问题的连续元素启发式二值法，提出了基于余弦相似度的特征选择方法。该文章将所提方法与六种先进算法在三个广泛使用的数据集上进行了比较，证明了该方法的优越性。Almasoudy 等人[29]提出了一种基于差分进化算法的入侵检测数据特征选择方法，该方法通过迭代地利用差分进化算法搜索最优特征子集并不断删除特征，使用 XGBoost 的计算精度评估这些特征，直到找到满足最高精度的最小特征。研究表明，该方法在 NSL-KDD 数据集的五分类和二分类中表现出较高的检测率和较低的误报率。Bonab 等人[30]指出高效的特征选择方法可以有效降低入侵检测系统的执行负载和存储复杂度，并指出结合包装器方法可以有效提高特征选择过程的准确性和效率。该文章提出了一种基于果蝇算法（FFA）和蚁狮优化算法（ALO）的混合包装特征选择算法，并在三个公开数据集：KDD CUP99、KSLKDD 和 UNSW-NB15 上证明了该方法的有效性。这些研究采用包装法，尽管实现了较好的结果，但通常需要大量迭代以搜索最佳特征组合，且未考虑在不平衡数据的网络入侵检测中应用包装法时适应度函数的改进。

Nazir 等人[31]提出了一种基于包装器的特征选择方法，称为“禁忌搜索—随机森林”。禁忌搜索是元启发式优化算法之一，用于特征子集的搜索。随机森林是此处用作学习算法的集成分类器之一。在该文章，原始数据集的特征减少了高达 60%，从而降低了模型的计算效率。受人工神经网络本质的启发，该文章提出了一种新的适应度函数作为基于包装的特征选择方法中的目标函数，旨在实现高检测率、高精度、低误报率。该文章利用三种机器学习和一种深度学习分类器（朴素贝叶斯、随机森林、K 最近邻和多层感知器）来测试所提出模型的性能。尽管该文章改进了适应度函数，但并没有考虑到不平衡因素的存在，并且 Tabu 搜索中的许多超参数设置需要花费大量的精力才能完美调整并获得最佳结果。因此，这是一种计算效率低下且不适合不平衡数据问题的方法。

为进一步提升特征选择的性能，部分学者尝试将不同类型的特征选择方法相结合，综合利用各方法的优势进行特征选择。此类融合方法旨在兼顾处理速度、算法复杂度以及特征选择的准确性，从而构建更加稳健且高效的模型。Zhao 等人[32]提出了一种基于相关性的特征选择与差分进化算法（CFS-DE），该方法利用 CFS 的相关性度量作为适应度函数来评估生成的新特征子集，并采用差分进化算法搜索最优特征子集。CFS 能够同时考虑特征与标签类之间及特征之间的关系，在保留重要特征的同时减少冗余特征的影响。最终，采用加权 Stacking 算法进行分类。该方法将过滤法与包装法相结合，为包装法提供了更为科学的搜索方向。Yin 等人[33]则提出了一种结合多层感知器网络的混合特征选择策略 IGRF-RFE，以应对多类网络异常检测问题。在初始阶段，该方法通过信息增益与随机森林的结合，有效缩小了特征子集的搜索空间。随机森林在此过程中优化了信息增益所选特征中高频但相对不重要的特征影响，确保了更多相关特征的纳入。在后续阶段，该方法采用基于机器学习的包装策略，通过递归特征消除进一步减少特征维度，并兼顾特征间的相关性。此方法成功融合了过滤法的高效性与包装法的相关性搜索优势。尽管这些方法结合了多种特征选择技术并实现了性能提升，但仍存在包装法依赖的模型易于过拟合、搜索效率较低及初始特征集随机性影响性能等问题。

5 问题和挑战

本研究全面总结了网络入侵检测领域的类别平衡方法和特征选择方法，问题表示的目的已被明确说明，并详细总结分析了多种相关方法的文献综述。虽然类别平衡方法和特征选择方法在提升网络入侵检测系统预测性能的应用上已经相当成功，但仍然存在一些问题，这些问题将在下面的章节中讨论。

一个真实的网络入侵数据集可能包含成百上千甚至更多的特征，而攻击类别会随着科技的进步逐渐出现多元化、混淆化的趋势。由于类别平衡和特征选择挑战涉及的数据集很大，因此发展的、使用的技术必须是可扩展的。如果一个类别平衡方法生成的合成样本与真实样本点密度分布趋于相同，或者一个特征选择算法在多个数据集中始终选择相同的特征子集，则认为它们是稳定的。

高维数据和不平衡数据集、所设计算法的可解释性、技术评估指标的稳定性以及算法可扩展性和泛化性能，这都是网络入侵检测系统所需要考虑的问题，研究人员可以在各种应用中提高类别平衡和特征选择优化技术的有效性和性能。基于此，本研究提出以下几点建议：

(1) 大多数时候，在训练过程中类别平衡和特征选择的标准往往和分类器的预测性能联系起来。因此，对于损失函数的设计不应只考虑分类器的准确率，而应该将特征性质、整体性能和局部表现综合起来考虑。对于特征选择方法的选择，可以将不同种类的方法结合起来，如进行过滤-包装两阶段的选择，从而改进搜索策略、传递函数和损失函数，使得最终的特征选择方法更具有普适性和高效性。

(2) 所使用或所设计的算法应该是能被用户可解释和可理解的。然而，类别平衡中的生成模型和特征选择中的一些元启发式优化技术可能会产生难以解释的复杂模型。研究人员应该研究可以产生可解释模型的技术，如决策树和线性模型。

(3) 类别平衡和特征选择都可能会产生不稳定的结果，

特别是当数据集很小或有噪声时。研究人员可以通过使用交叉验证评估他们的方法的稳定性。其次，基于深度学习的类别平衡方法或基于启发式的特征选择方法在计算上很昂贵，研究人员需要评估他们的方法的可扩展性，并研究可能提高效率的并行化策略。

(4) 在某些实际的网络入侵检测系统设计过程中，类别平衡和特征选择的优化可能需要考虑多个目标。研究人员应该研究多目标优化技术，可以处理多个相互冲突的目标。

总之，类别平衡和特征选择领域的研究人员应该考虑上述建议，以解决他们研究中的问题和挑战。解决这些挑战有助于提高类别平衡和特征选择优化技术的效率和有效性。

6 总结

本研究提供了一个全面详细的网络入侵检测中类别平衡与特征选择方法的应用的文献综述，还提供了主流类别平衡和特征选择方法的详细描述和数学模型，这有助于研究人员充分理解这个问题。在文献综述部分，本研究详细讨论了类别平衡和特征选择方法在入侵检测领域的应用，并分析这些文献所使用方法的优势与缺陷；在方法概述部分，本研究描述了多种主流方法的的基本定义，意义和分类，并以表格形式提供直观的性能对比。

对于类别平衡方法的应用，近年来研究人员逐渐将深度学习技术应用于类别平衡任务中，其中最具代表性的是基于 GAN 的系列模型，相比于传统的统计方法，深度学习方法能够生成更真实更符号真实样本趋势的样本，但与之出现的挑战是模型收敛问题和参数调优问题；对于特征选择方法的应用，以启发式算法为代表的包装法开始被广泛应用，启发式算法参数初始化、搜索策略的改进、传递函数的使用以及损失函数的设计上都具有了快速的发展。诚然，现阶段的研究人员逐渐聚焦于多种方法相结合的应用，比如先对数据集使用过滤法筛选排名，再使用包装法顺序前向搜索，从而集成不同种类方法的优势。

在文献综述分析结束后，本研究针对总结的文献的特点进行总结，指出了研究人员在网络入侵检测领域应解决的问题和迎接的挑战，并给出具体的建议和展望未来研究方向。

参考文献

[1] 杨秀璋,彭国军,刘思德,等.面向 APT 攻击的溯源和推理研究综述[J/OL].软件学报,1-50[2024-12-05].

[2] Langner R. Stuxnet: Dissecting a cyberwarfare weapon[J]. IEEE Security & Privacy, 2011, 9(3): 49-51.

[3] Antiy Laboratory. A comprehensive analysis report o n Ukraine power grid outage[R/OL]. (2016)[2024-12-09].https://www.antiy.com/response/A_Comprehensive_Analysis_Report_on_Ukraine_Power_Grid_Outage/A_Comprehensive_Analysis_Report_on_Ukraine_Power_Grid_Outage.html.

[4] FireEye. Highly evasive attacker leverages SolarWinds supply chain to compromise multiple global victims with SUNBURST backdoor[EB/OL]. (2020)[2024-12-09]. https://www.fireeye.com/blog/threat-research/2020/12/evasive-attacker-leverages-solarwinds-supply-chain-compromises-with-sunburst-backdoor.html.

[5] Digital Security Technology Group. Investigative repo rt on northwestern Polytechnical University’s discovery of US NSA cyber attack[R/OL]. (2022)[2024-12-09]. https://mp.weixin.qq.com/s/0ReOzQMM5GS4xXRUPpKCvA.

[6] Alshamrani A, Myneni S, Chowdhary A, et al. A survey on advanced persistent threats: Techniques, solutions, challenges, and research opportunities[J]. IEEE Communications Surveys & Tutorials, 2019, 21(2): 1851-1877.

[7] Al S, Dener M. STL-HDL: A new hybrid network intrusion detection system for imbalanced dataset on big data environment[J]. Computers & Security, 2021, 110: 102435.

[8] Hammad M, Hewahi N, Elmedany W. MMM-RF: A novel high accuracy multinomial mixture model for network intrusion detection systems[J]. Computers & Security, 2022, 120: 102777.

[9] Liu J, Gao Y, Hu F. A fast network intrusion detection system using adaptive synthetic oversampling and LightGBM[J]. Computers & Security, 2021, 106: 102289.

[10] Dunmore A, Jang-Jaccard J, Sabrina F, et al. A comprehensive survey of generative adversarial networks (gans) in cybersecurity intrusion detection[J]. IEEE Access, 2023.

[11] Arjovsky M, Chintala S, Bottou L. Wasserstein generative adversarial networks[C] // International Conference on Machine Learning. PMLR, 2017: 214-223.

[12] Gulrajani I, Ahmed F, Arjovsky M, et al. Improved training of wasserstein gans[C] // Advances in Neural Information Processing Systems. 2017, 30.

[13] Sánchez-Maroño N, Alonso-Betanzos A, Tombilla-Sanromán M. TOMBILLA-SAÑROMÁN M. Filter methods for feature selection[C] // Lecture Notes in Computer Science. 2007, 4881: 178-187.

[14] Zouhri H, Idri A, Ratnani A. Evaluating the impact of filter-based feature selection in intrusion detection systems[J]. International Journal of Information Security,

2024, 23: 759-785.

[15] Lal T N, Chapelle O, Weston J, et al. Embedded methods[M] // Feature Extraction: Foundations and Applications. Berlin, Heidelberg: Springer, 2006: 137-165.

[16] Liu Z, Thapa N, Shaver A, et al. Using embedded feature selection and CNN for classification on CCD-INID-V1—a new IoT dataset[J]. Sensors, 2021, 21(14): 4834.

[17] Weston S, Mukherjee S, Chapelle O et al. Feature selection for SVMs[C] // Advances in Neural Information Processing Systems. 2000, 12: 526-532.

[18] Bajer D, Zonč B, Dudjak M, et al. Performance analysis of SMOTE-based oversampling techniques when dealing with data imbalance[C] // 2019 International Conference on Systems, Signals and Image Processing (IWSSIP). New York: IEEE, 2019: 265-271.

[19] Md Alamin Talukder, Khondokar Fida Hasan, Md Manowarul Islam, et al. A dependable hybrid machine learning model for network intrusion detection[J]. Journal of Information Security and Applications, 2023, 72: 103405.

[20] Liu Jingmei, Gao Yuanbo, Hu Fengjie. A fast network intrusion detection system using adaptive synthetic oversampling and LightGBM[J]. Computers & Security, 2021, 106: 102289.

[21] Fu Y, Du Y, Cao Z, et al. A Deep Learning Model for Network Intrusion Detection with Imbalanced Data[J]. Electronics, 2022, 11(6): 898.

[22] Kim K. GAN based augmentation for improving anomaly detection accuracy in host-based intrusion detection systems[J]. International Journal of Engineering Research and Technology, 2020, 13(11): 3987.

[23] Lee J, & Park K. GAN-based imbalanced data intrusion detection system[J]. Personal and Ubiquitous Computing, 2021, 25: 121-128.

[24] Xiong W.D, Luo K.L, Li R. AIDTF: Adversarial training framework for network intrusion detection[J]. Computers & Security, 2023, 103141.

[25] Kumar, V., Sinha, D., 2023. Synthetic attack data generation model applying generative adversarial network for intrusion detection. Comp. Secur. 125, 103054.

[26] Srivastava A, Sinha, D, Kumar, V. WCGAN-GP based synthetic attack data generation with GA based feature selection for IDS[J]. Computers & Security, 2023, 134: 103432.

[27] Raileanu L.E, Stoffel K. Theoretical comparison between the Gini index and information gain criteria[J]. Annals of Mathematics and Artificial Intelligence, 2004, 41(1): 77-93.

[28] Alazzam H, Sharieh A, Sabri, K.E. A feature selection algorithm for intrusion detection system based on pigeon inspired optimizer[J]. Expert Systems with Applications, 2020, 148: 113249.

[29] Almasoudy F.H, Al-Yaseen W.L, Idrees A.K. Differential evolution wrapper feature selection for intrusion detection system[J]. Procedia Computer Science, 2020, 167: 1230-1239.

[30] Bonab M, Ghaffari A, Gharehchopogh F, et al. A wrapper-based feature selection for improving performance of intrusion detection systems[J]. International Journal of Communication Systems, 2020, 33(12): e4434.

[31] Nazir A, Khan R. A novel combinatorial optimization-based feature selection method for network intrusion detection[J]. Computers & Security, 2021, 102: 102164.

[32] Zhao R, Mu Y, Zou L, et al. A hybrid intrusion detection system based on feature selection and weighted stacking classifier[J]. IEEE Access, 2022, 10: 71414-71426.

[33] Yin Y, Jang-Jaccard J, Xu W, et al. IGRF-RFE: A hybrid feature selection method for MLP-based network intrusion detection on UNSW-NB15 dataset[J]. Journal of Big Data, 2023, 10(1): 15.