文章编号：2096-1618(2025)02-0125-07

基于惩罚回归的高噪声流量分类

白凯毅[1]， 盛志伟[1,2]， 黄源源[1,2]

(1. 成都信息工程大学网络空间安全学院，四川 成都 610225；2. 先进密码技术与系统安全四川省重点实验室，四川 成都 610225)

摘要：针对网络流量数据容易受到干扰的现实情况，引入带噪声标签学习的思想，并人为添加噪声以模糊化特征。先建立特征和标签之间的线性关系，然后用 mean-shift 参数识别噪声数据。通过人工添加对称噪声和非对称噪声模拟现实情况下的各种干扰信息。由此提出一个基于 L2 正则的高噪声流量分类模型(PR-2)，通过将流量转换为图像并应用 L2 正则化方法来处理带噪声的标签，以提高高噪声流量下分类模型的性能。在 USTC-TF2016 数据集上验证了本方法的有效性，并与 LSTM、BiTCN、BoAu、CL、INCV、FINE 方法进行对比。实验结果表明，PR-2 方法在对称噪声和非对称噪声的噪声比为0.8的情况下仍能取得95.16% 和86.15% 的准确率，证明其在处理高噪声数据方面的有效性和可用性。

关 键 词：高噪声流量；流量分类；深度学习；带噪声标签学习

中图分类号：TP183 **文献标志码**：A

doi：10.16836/j.cnki.jcuit.2025.02.001

0 引言

随着互联网和通信技术的快速发展，网络流量分类对于高效实现网络管理、优化资源分配、保证服务质量有非常重要的作用[1]。在过去十几年中，流量分类领域的进步大多依赖于机器学习和深度学习。这其中的模型训练通常需要大量有准备标签的数据，以获得良好的泛化能力和分类能力。然而，在将捕获的流量作为数据集进行处理时，由于在捕获流量过程中，通信信道可能出现干扰信息或者网络传感器和嗅探工具的缺陷，使捕获的流量受到干扰，产生噪声、异常行为或冗余数据，这些都可能干扰标签的正确性和可靠性。因此，数据质量成为一个亟待解决的难题[2]。

在现今的分类模型中，噪声数据是影响分类性能的一个重要因素。它们通过引入误导性的信息，使分类器难以准确地学习和泛化。当分类模型在训练过程中接触到噪声数据时，就会导致模型过度拟合或错误地调整决策边界，从而降低模型的性能和准确度。而带噪声标签学习(LNL)[3-4]则能更好地解决这一问题，可以在保留噪声标签的同时进行训练，通过调整模型或损失函数使其能够在噪声数据存在的情况下进行学习和预测。这就使 LNL 在大规模数据集和无法全面清洗数据的情况下更具可行性，在图像领域证明了它的应用价值。

本文旨在解决流量分类中的高噪声流量数据问题，并提出一种创新方法——PR-2。首先把流量数据集的特征提取为灰度图，然后结合图像领域的优秀方法，提出一种理论上有保证的噪声标签检测框架来检测噪声数据。并从统计学的角度出发，在有理论支持的情况下来识别噪声数据。用 mean-shift 参数识别噪声数据，结合半监督学习对噪声数据进行处理，并通过人为添加的对称噪声和非对称噪声来模拟现实环境中各种干扰信息的影响。

为验证 PR-2 方法的有效性，在 USTC-TF2016 上进行实验，并与 LSTM[5]、BiTCN[6]、BoAu[7]、CL[8]、INCV[9]、FINE[10]等方法进行对比。实验结果表明，在对称噪声和非对称噪声的高噪声情况下，PR-2 方法均取得了较高的准确率，证明其在处理噪声数据方面的有效性和可用性。

1 相关工作

1.1 机器学习

传统的使用机器学习算法来进行流量分类的方法主要包括监督学习和无监督学习，大多准确率低、分类效率低且易受到噪声的干扰。

Anderson 等[11]比较了线性回归、逻辑回归、决策

收稿日期：2023-09-20

基金项目：国家重点研发计划资助项目(2022YFB3103103)；四川省重点研发计划资助项目(2022YFS0571)；四川网络文化研究中心资助项目(WLWH22-18)；四川省自然科学基金资助项目(2022NSFSC0557)

通信作者：盛志伟. E-mail：7782988@qq.com

树、随机森林、支持向量机和多层感知器这 6 种机器学习方法在噪声数据1.5% ~5% 的情况下的分类性能，发现它们的性能都随着时间的推移而下降。

Gethami 等[12]研究了噪声对基于 ML 的 IDS 精度的影响程度，使用的机器学习算法是决策树、随机森林、支持向量机、人工神经网络和朴素贝叶斯。然后将不同级别的噪声（5%、10%、20% 和 30%）添加到每个训练数据集中，用这几种算法进行测试。发现算法都随着噪声的增加性能逐渐下降，RF 的下降程度尤为明显。

Yuan 等[13]提出一种基于无监督学习的高噪声流量清洗的算法。该算法抛开标签的束缚，通过使用自动编码器进行低维表示和置信度评估来识别和拒绝错误样本，并尽可能地保留硬样本，以提高分类器的泛化能力。但是它的局限性明显，单纯地通过数据清洗来识别和剔除错误样本，会导致一些真实的硬样本被错误地标记为错误样本并被丢弃，从而影响分类器的泛化性能，特别是当这些硬样本在测试集中出现时。

Yuan 等[7]提出一种基于边界增强的噪声标签恶意流量检测（BoAu），使用 K-means 作为 BoAu 的无监督分类器。BoAu 在训练过程中使用所有样本（包括所有硬样本）以构建更准确的决策边界，使模型保持较高的噪声容忍度。但它很依赖无监督聚类算法（K-means）作为分类器，这种简单的聚类算法无法充分利用复杂数据集中的潜在信息。

1.2 深度学习

传统机器学习的分类性能严重依赖手工提取网络流量的特征，不仅效率低，而且很容易受到干扰。后来开始结合深度学习算法对网络流量进行分类，利用深度学习可以自动学习特征表示和处理复杂数据，提高了流量的分类性能。但是，噪声依然会对分类造成很大的影响。Zhang 等[14]研究发现，在存在大量噪声的流量数据中，DNN 可能会过度适应这些噪声，而无法正确地泛化到真实的流量样本上。由于 DNN 过于依赖训练数据中的损坏标签进行模型训练，因此可能无法正确地推广到新的、未见过的测试数据。Hwang 等[5]把词嵌入和 LSTM 结合在一起，虽然可以提高流量的分类性能，但是加入噪声后的分类效果却大幅度降低。Chen 等[6]提出一种基于双向时间卷积网络的异常网络流量检测模型 BiTCN，使用时间卷积网络（TCN）来捕捉网络流量的序列特征，引入指数线性单元（ELU）激活函数提高模型的表达能力和性能，并且把原有的单向模型改进为双向模型，使模型可以同时考虑正向和逆向的网络流量信息，从而提高检测模型对异常流量的敏感度和准确性。但是由于没有考虑噪声的影响，对其实验进行复现并加入噪声后发现，随着噪声比例的增加，性能逐渐降低，80% 噪声比例时，准确率降低约 20%。Fallah 等[15]将流量数据建模为一个流量序列，并采用基于序列的深度学习模型。通过添加一些随机的额外数据来模拟噪声，最终发现在噪声比例 50% 的情况下仍有 80% 的准确率。不过，它只模拟了一种噪声情况，而现实中产生噪声的方式有很多种，也有不同的噪声，并且噪声比例也不算高，因而检测高噪声情况就很难有效。由于噪声标签对深度神经网络的泛化能力有严重的负面影响，因此现在深度学习应用场景的任务之一就是从噪声标签中学习。这种方法广泛应用于计算机视觉、自然语言处理、语音识别等方面，带噪声标签学习也成为一个专门的研究领域。

1.3 带噪声标签学习

由于近几年深度学习的快速发展，对带噪声标签学习有一定的研究。Han 等[16]提出一个“Masking”的人工辅助方法，由此推导一个包含结构先验的结构感知概率模型，形成一个新的网络结构来提高识别效率。Lyu 等[17]通过修改损失函数，提出一种课程损失的方法，以自适应地选择样本进行模型训练，提高了模型的鲁棒性。Wang 等[18]通过添加 L1 正则的惩罚函数进行噪声处理，把一些特征化为零，但是运用在流量分类上就很容易把重要的特征消除。CL[8]、INCV[9] 和 FINE[10] 等数据清洗方法能有效使用监督分类器来分析数据中存在的标签错误，并从中获取适当的特征以进行评估，这意味着分类器在数据包含错误标签的情况下需要具备有效提取和学习特征的能力。

2 基于惩罚回归的高噪声流量数据分类

2.1 问题描述

在网络流量中，流量的特征和对应的标签总会存在线性关系：

$$\boldsymbol{y}_i = \boldsymbol{x}_i\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{1}$$

式中，流量的特征为 $\boldsymbol{y}$，$\boldsymbol{x}$ 是对应的标签输出为 one-hot 编码，$\boldsymbol{\beta}$ 是系数矩阵，$\boldsymbol{\varepsilon}$ 是随机的噪声。如果噪声存在，那么它的输出肯定是密集的。而判断这个数据是

不是噪声,可以检查它的误差 u, $\| u \|$ 越大,说明它是异常点的可能性越大。误差 u:

$$u_i = \boldsymbol{y}_i - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}^{\mathrm{T}} \tag{2}$$

用数学经典留一法[19]来检验外部学生化残差。

噪声数据通过回归模型中求解的非零的 mean-shift 参数 $\boldsymbol{\gamma}$ 来识别,则线性回归可以重新解释为

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\gamma} + \boldsymbol{\varepsilon} \tag{3}$$

$\boldsymbol{\gamma}$ 是用来识别和去除噪声数据的,求解则

$$\arg\min_{\boldsymbol{\beta},\boldsymbol{\gamma}} \frac{1}{2} \| \boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta} - \lambda \|_F^2 + \lambda_i \| \boldsymbol{\gamma}_i \|_2^2 \tag{4}$$

其中,λ 是正则化系数。为简化表示,记 $\tilde{\boldsymbol{Y}} = \boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}$,目标函数可以重写为

$$\arg\min_{\boldsymbol{\gamma}} \frac{1}{2} \| \tilde{\boldsymbol{Y}} - \boldsymbol{\gamma} \|_2^2 + \lambda \| \boldsymbol{\gamma}_i \|_2^2 \tag{5}$$

可以把目标函数展开:

$$\frac{1}{2}(\tilde{\boldsymbol{Y}} - \boldsymbol{\gamma})^{\mathrm{T}}(\tilde{\boldsymbol{Y}} - \boldsymbol{\gamma}) + \lambda \boldsymbol{\gamma}^{\mathrm{T}}\boldsymbol{\gamma} \tag{6}$$

需要求解的是使目标函数最小化的 $\boldsymbol{\gamma}$。为求解最小化问题,可以对目标函数关于 $\boldsymbol{\gamma}$ 求导,并令导数等于零:

$$\frac{\partial}{\partial \boldsymbol{\gamma}}\left(\frac{1}{2}(\tilde{\boldsymbol{Y}} - \boldsymbol{\gamma})^{\mathrm{T}}(\tilde{\boldsymbol{Y}} - \boldsymbol{\gamma}) + \lambda \boldsymbol{\gamma}^{\mathrm{T}}\boldsymbol{\gamma}\right) = 0 \tag{7}$$

将上述导数求解,可以得到:

$$\boldsymbol{\gamma} = \frac{\tilde{\boldsymbol{Y}}}{1 + 2\lambda} \tag{8}$$

它代表了对应于噪声或离群值的估计 λ 的解。由于 L2 正则化的存在,解 λ 不再是稀疏的,而是具有更均衡的权重分布。

考虑一个包含图片-标签对 $(I_i, y_i)_{i=1}^n$ 的数据集,其中图片 $I_i \in \mathcal{I} \subseteq \mathbb{R}^m$,对应的标签为 $y_i \in \mathcal{C} \subseteq \mathbb{R}$,且类别集合$\mathcal{C}$的大小为 c。假设每个实例的标签 y_i 都是从真实标签 y_i^* 经过未知破坏过程得到的,那么目标就是通过一个神经网络模型,由分类器$g(\cdot)$和特征提取器$f(\cdot)$组成,对任意输入图片 $I \in \mathcal{I}$ 预测其对应的基础真实标签 $y^* \in \mathcal{C}$。通常,网络首先将输入图片 I_i 编码为 $x_i = f(I_i)$,然后利用分类器$g(\cdot)$产生对应的 soft-max 概率预测:

$$\hat{\boldsymbol{y}}_i = \boldsymbol{g}(\boldsymbol{x}_i) \tag{9}$$

目标是最小化损失函数:

$$\boldsymbol{\varphi}(\boldsymbol{\beta},\boldsymbol{\gamma}) = \frac{1}{2} \| \boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{\gamma} \|_F^2 + \sum_{i=1}^{n} \boldsymbol{\lambda}_i \| \gamma_i \|_2^2 \tag{10}$$

其中,$\boldsymbol{X}$ 表示特征矩阵,$\boldsymbol{Y}$ 表示标签矩阵,$\boldsymbol{\beta}$ 表示线性回归模型的系数矩阵,$\boldsymbol{\gamma}$ 表示 mean-shift 参数向量,λ_i 表示正则化系数,n 表示样本数量。

目标是找到最优的 $\boldsymbol{\beta}$ 和 $\boldsymbol{\gamma}$,使损失函数最小化。可以通过求解以下最优化问题来实现:

$$\arg\min_{\boldsymbol{\beta},\boldsymbol{\gamma}} \boldsymbol{\varphi}(\boldsymbol{\beta},\boldsymbol{\gamma}). \tag{11}$$

为求解该最优化问题,可以采用梯度下降进行迭代优化。首先,计算损失函数关于 $\boldsymbol{\beta}$ 和 $\boldsymbol{\gamma}$ 的梯度:

$$\frac{\partial \boldsymbol{\varphi}}{\partial \boldsymbol{\beta}} = -\boldsymbol{X}^{\mathrm{T}}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{\gamma}) \tag{12}$$

$$\frac{\partial \boldsymbol{\varphi}}{\partial \boldsymbol{\gamma}} = -(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{\gamma}) - 2\sum_{i=1}^{n} \lambda_i \boldsymbol{\gamma}_i \tag{13}$$

然后,使用梯度下降更新 $\boldsymbol{\beta}$ 和 $\boldsymbol{\gamma}$ 的值

$$\boldsymbol{\beta} \leftarrow \boldsymbol{\beta} - \eta \frac{\partial \boldsymbol{\varphi}}{\partial \boldsymbol{\beta}} \tag{14}$$

$$\boldsymbol{\gamma} \leftarrow \boldsymbol{\gamma} - \eta \frac{\partial \boldsymbol{\varphi}}{\partial \boldsymbol{\gamma}} \tag{15}$$

其中 η 表示学习率或步长参数,用于控制优化过程的更新速度。通过多次迭代更新 $\boldsymbol{\beta}$ 和 $\boldsymbol{\gamma}$,直到达到收敛条件或达到最大迭代次数,可以得到最优的 $\boldsymbol{\beta}$ 和 $\boldsymbol{\gamma}$ 的估计值。最后,根据得到的 $\boldsymbol{\gamma}$ 值,可以将对应的实例识别为噪声数据,因为非零的 $\boldsymbol{\gamma}_i$ 对应的实例可能受到噪声的影响。

Huber's M-estimate[20] 是一种稳健的回归方法。本文使用 Huber´s M-estimate 来获得对数据中极端观测值或异常值不那么敏感的参数估计。

$$\arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \rho\left(\frac{\boldsymbol{y}_i - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}}{\sigma}; \lambda\right) + \frac{1}{2cn\sigma} \| \boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta} \|_F^2 \tag{16}$$

把 λ 看作是噪声数据的一个指标,其中 $\| \boldsymbol{\gamma}_i \|$ 越大意味着实例 i 经历了更多的噪声破坏,集合 n 定义为噪声的样本集 $n = i: \| \boldsymbol{\gamma}_i \| \neq 0$,然后把 $\boldsymbol{\gamma}$ 带入式(6)求解

$$\arg\min_{\boldsymbol{\gamma}} \frac{1}{2} | \tilde{\boldsymbol{Y}} - \tilde{\boldsymbol{X}}\boldsymbol{\gamma} |_2^2 + \lambda_i \| \boldsymbol{\gamma}_i \|_2^2 \tag{17}$$

其中 $|\cdot|_2$ 表示 L2 范数,通过最小化这个目标求解,然后采用分块递归算法求解 $\boldsymbol{\gamma}$,随着 $\boldsymbol{\gamma}$ 逐渐减小,惩罚项的影响逐渐降低,这个效果主要是由模型选择的。由于选择越早就越有可能存在噪声,因此把所有的样本按照它选择时间的下降顺序排序,定义为 $T_i = \sup\{\lambda: (\boldsymbol{\gamma}_i) \neq 0\}$,$T_i$ 越大,认定它越有可能是噪声样本。

2.2　含噪声样本的学习

监督训练学习是在检测到噪声集合 n 时,把噪声数据从训练数据集中去掉,用剩余干净的数据训练网络模型。假设通过 softmax 预测的结果是 $\boldsymbol{X}\boldsymbol{W}_f$,$\boldsymbol{W}_f$ 是最后一层全连接层的权重,但是在式(3)中假设的是两者线性相关。为减少差距,在交叉熵损失后面添加一个 q 惩罚来鼓励两者的线性关系。

$$\varphi(x_i,y_i)=1_{i\notin n}(\varphi_{CE}(x_i,y_i)+\lambda\|x_i^{\mathrm{T}}W_f\|_2) \quad (18)$$

其中,λ 是正则化参数,$1_{1_{\notin n}}$是指示函数,只有在干净数据上计算损失,这个正则项鼓励近似称为一个 one-hot 编码向量。

半监督学习是通过在干净数据和噪声数据中插值来生成部分图像,然后使用这些插值数据进行网络的训练。具体是通过以下公式生成插值数据:

$$\tilde{x}=M\odot x_{\text{干净}}+(1-M)\odot x_{\text{噪声}}$$
$$\tilde{y}=\lambda y_{\text{干净}}+(1-\lambda)y_{\text{噪声}} \quad (19)$$

然后使用插值后的网络来训练:

$$\varphi(\tilde{x},\tilde{y})=\varphi_{\mathrm{CE}}(\tilde{x},\tilde{y}) \quad (20)$$

由于 $\tilde{y}$ 是插值得到的,不再是 one-hot 向量,因此在使用插值数据进行训练时可以不使用 L2 惩罚项。

2.3 噪声集恢复

这一步是要减少噪声数据对模型性能的负面影响。通过噪声集恢复来识别和剔除噪声数据,提高模型性能的鲁棒性。通过引入正则化参数,并结合上述的其他条件可以恢复噪声集。通过使用岭回归(Ridge Regression)的方式来恢复噪声集,在目标函数中添加 L2 惩罚性,利用如下函数:

$$\arg\min_{\gamma}\frac{1}{2}\|\tilde{y}-\mathring{X}\tilde{\gamma}\|_2^2+\lambda\|\tilde{\gamma}\|_2^2 \quad (21)$$

通过最小化目标函数来恢复、剔除噪声数据。

3 实验结果与分析

3.1 数据集

本文使用 USTC-TFC2016 数据集[21],由中国科技大学发布,是一个公开的网络流量数据集,包含了正常流量和恶意流量,涵盖 8 类常见的应用程序。正常的流量包括 HTTP、FTP、SMTP、POP3 等;恶意流量包括 Cridex、Shifu、Miuref、Geodo 等 10 种不同类型的恶意流量。该数据集常被用于网络流量分类任务,其详细信息如表 1 所示。

表 1 USTC-TFC2016 数据集的详细信息

数据集类型	流量来源	流量类型	总数
正常流量	File Transfer Protocol	Data Trafnsfer	358903
	Bit Torrent	P2P	15230
	Facetime	Video	6010
	MySQL	Database	201020
	Outlook	Mail	15200
	Skype	Chat	12200
	Server message block	Data Trafnsfer	926352
	Gmail	Mail	25044
	Weibo	Social NetWork	2609260
	World of Warcraft	Game	135900
恶意流量	Htbot	Malicious machine traffic	168981
	Shifu	Banking Trojans	500077
	Miuref	Trojans	82408
	Tinba	Banking Trojans	22012
	Neris	Spam	494068
	Geodo	Botnet	224258
	Zeus	Trojans	87038
	Nsis-ay	Botnet	349987
	Virut	Virus	440649
	Gridex	Banking Trojans	459952

3.2 实验过程及结果分析

本文选用的深度学习模型是两层卷积层和两层全连接,然后输入到 PR-2 模型中进行进一步的训练,实验流程见图 1。

本文对比了7种方法,即:LSTM、BiTCN、BoAu、CL、INCV、FINE以及本文提出的PR-2在不同噪声比例下的分类性能,结果见表 2。其中,对称噪声和非对称噪声的噪声比例都为0.2、0.4、0.6、0.8。

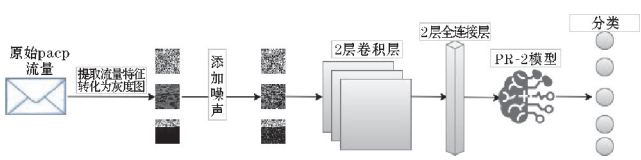


图 1 模型流程图

表 2　USTC-TF2016 数据集上 7 种方法在对称和非对称噪声的各个噪声比例情况下的准确度对比　　单位:%

方法	对称噪声噪声比例				非对称噪声噪声比例			
	0.2	0.4	0.6	0.8	0.2	0.4	0.6	0.8
LSTM	90.67	82.33	74.14	61.36	87.13	83.53	70.94	60.08
BiTCN	92.32	85.28	69.17	57.15	88.58	81.83	72.73	58.12
BoAu	98.54	97.46	95.24	92.64	98.98	92.02	75.41	62.15
CL	90.95	83.26	79.23	70.42	89.92	80.41	74.87	62.73
INCV	90.35	84.87	80.11	74.30	88.91	81.07	71.13	60.78
FINE	82.86	73.73	65.84	62.46	80.82	75.45	68.17	60.21
PR-2	98.98	98.73	97.93	95.16	99.02	98.72	89.48	86.15

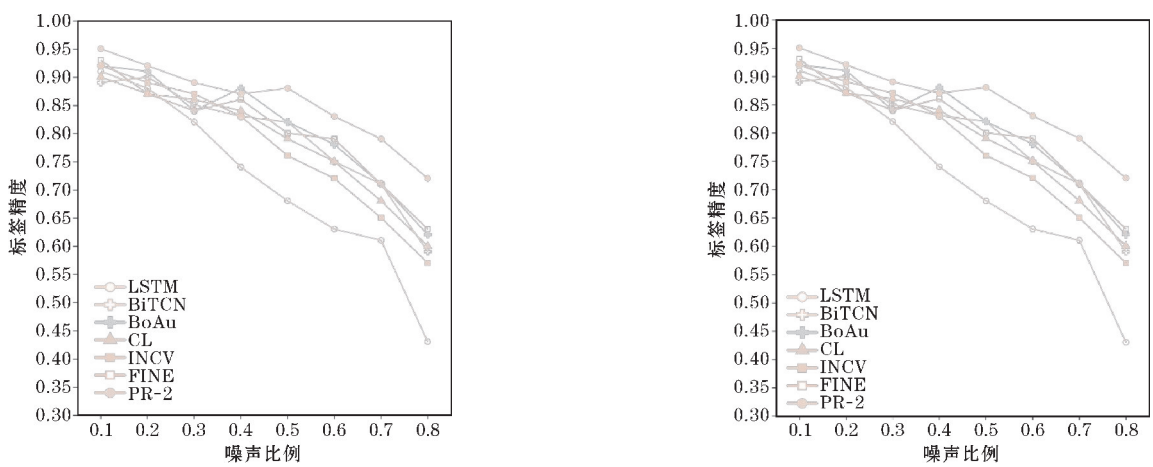
由表 2 可知,在对称噪声比例下,所有的分类方法准确率都有所下降。这是因为引入了更多的噪声标签,导致学习算法在训练过程中受到错误标签的干扰,从而降低了分类性能。在对称噪声比例为0.2的时候,PR-2 方法表现得更好,达到了98.98% 的准确率,略高于其他方法。

在非对称噪声比例为0.2、0.4、0.6,0.8时,所有方法的正确率都高于对称噪声比例下的对应数值。这是因为非对称噪声相比对称噪声而言,学习算法更容易通过一致的真实标签来学习,从而提高了分类性能。在非对称噪声比例为0.2时,PR-2 方法具有最高的准确率,为99.02% 。

实验结果的数据分析表明,在 USTC-TF2016 数据集上,本文提出的 PR-2 方法相对于其他 6 种方法在不同噪声比例下都表现出较好的分类性能。尤其是在对称噪声比例为0.2和非对称噪声比例为 0.2 的情况下,PR-2 方法具有最高的分类准确率。并且在噪声比例为0.8的情况下,也有95.16% 和86.15% 的分类性能。

3.3　标签精度

除了准确性外,评判样本选择算法能力指标还有标签的精度。从图 2 可以看出,随着噪声比例的不断增加,各个算法标签精度明显呈下降趋势。而且在噪声比例0.1 ~0.4,所有算法都有较强的标签精度,但是大于0.4之后,其他的算法都开始大幅度下降,而本文提出的 PR-2 算法则仍然保持相对稳定的标签精度。



(a)对称噪声　　(b)非对称噪声

图 2　对称噪声和非对称噪声下的标签精度

3.4　消融实验

为评估本文提出的框架中各种模块的有效性,使用对称噪声率为 20% 的 USTC-TF2016 数据集进行消融研究,旨在验证每个模块对整体性能的影响,对不同的组件组合进行评估,结果如表 3 所示。

表 3　PR-2 使用不同模块的精度

模型	准确度/%
CE	82.8
CE+PR-2	88.7
CE+q	86.2
CE+PR-2+q	98.98

表 3 中“CE”代表的是使用交叉熵损失法的基线方法,准确率达到了82.8%。本文采用更先进的配置:

“CE + PR-2”:只对 PR-2 模块识别出的干净数据使用交叉熵损失,准确率大幅提高到88.7%。

“CE + q”:对所有训练数据都采用式(18),准确率为86.2%。

在进一步分析的基础上,将各种组件结合起来,探索它们的协同效应:

“CE + q + PR-2”:整合式(18)方法、PR-2 模块和交叉熵损失。在所有配置中,该综合模型的准确率最高,达到98.98%。

结果表明,与标准的交叉熵损失方法相比,本文的噪声检测框架能显著提高性能。此外,基于 PR-2 的框架中所有相关组件的组合带来了最佳的准确性。

3.5 q 范数的影响

不同 q 下 PR-2 的准确度见图 3,研究 q 范数对本文框架的影响,通过对从0.05到 1 的一系列 q 值进行实验,发现 q 值对本文模型性能的细微影响。

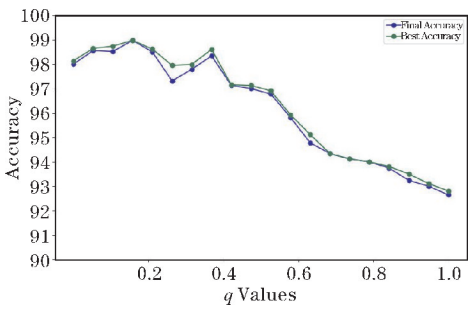


图 3 不同 q 下 PR-2 的准确度

综上所述,q 值的选择对本文框架的行为有显著的影响。较小的 q 值会促进线性关系的形成,这与本文提出的方法的预期行为一致。值得注意的是,过小的 q 值会损害网络的表示能力,因此,出现一条明显的凸精度曲线,这表明必须取得最佳平衡。

在所考察的一系列 q 值中,准确度曲线呈现出一个最佳性能区域。值得注意的是,在本文实验范围内,0.2 的 q 值是最有效的选择。对实验结果的综合分析证实了这一结论,并选择 $q=0.2$作为本文框架的最佳配置。

4 结束语

本研究聚焦于高噪声流量分类领域,旨在解决现实中流量数据收集时受到干扰的挑战。在这个领域中,深度学习模型在识别流量类型时通常依赖于强大而干净的数据集。然而,在实际场景中,流量数据常常会受到各种干扰,导致标签数据含有噪声。为应对这一问题,本文将流量数据视为一个数学模型,并引入惩罚回归的方法来处理噪声数据。实验结果表明,在对称和非对称流量噪声比例高达0.8的情况下,PR-2 模型显示出显著的分类性能,为高噪声流量数据的分类提供了有效的解决方案。未来的研究可以进一步探索 PR-2 模型在其他数据集和实际场景中的适用性,并与其他先进的噪声检测和分类方法进行比较。

致谢:感谢成都市科技项目(2023-XT00-00002-GX)对本文的资助

参考文献:

[1] Azab A, Khasawneh M, Alrabaee S, et al. Network traffic classification: Techniques, datasets, and challenges[J]. Digital Communications and Networks, 2024, 10(3): 676-692.

[2] Guerra J L, Catania C, Veas E. Datasets are not enough: Challenges in labeling network traffic[J]. Computers & Security, 2022, 120: 102810.

[3] Song H, Kim M, Park D, et al. Learning from noisy labels with deep neural networks: A survey[J]. IEEE Transactions on Neural Networks and Learning Systems, 2023, 34(11): 8135-8153.

[4] Nigam N, Dutta T, Gupta H P. Impact of noisy labels in learning techniques: a survey[C]. Advances in Data and Information Sciences: Proceedings of ICDIS 2019. Springer Singapore, 2020: 403-411.

[5] Hwang R H, Peng M C, Nguyen V L, et al. An LSTM-based deep learning approach for classifying malicious traffic at the packet level[J]. Applied Sciences, 2019, 9(16): 3414.

[6] Chen J, Lv T, Cai S, et al. A novel detection model for abnormal network traffic based on bidirectional temporal convolutional network[J]. Information and Software Technology, 2023, 157: 107166.

[7] Yuan Q, Liu C, Yu W, et al. BoAu: Malicious traffic detection with noise labels based on boundary augmentation[J]. Computers & Security, 2023, 131: 103300.

[8] Northcutt C, Jiang L, Chuang I. Confident learning: Estimating uncertainty in dataset labels[J]. Journal of Artificial Intelligence Research, 2021, 70: 1373-1411.

[9] Chen P, Liao B B, Chen G, et al. Understanding and utilizing deep neural networks trained with

noisy labels[C]. International Conference on Machine Learning. PMLR,2019:1062-1070.

[10] Kim T,Ko J,Choi J H,et al. Fine samples for learning with noisy labels[J]. Advances in Neural Information Processing Systems,2021,34:24137-24149.

[11] Anderson B,McGrew D. Machine learning for encrypted malware traffic classification:accounting for noisy labels and non-stationarity[C]. Proceedings of the 23rd ACM SIGKDD International Conference on knowledge discovery and data mining. 2017:1723-1732.

[12] Al-Gethami K M,Al-Akhras M T,Alawairdhi M. Empirical evaluation of noise influence on supervised machine learning algorithms using intrusion detection datasets[J]. Security and Communication Networks,2021,2021(1):8836057.

[13] Yuan Q,Zhu Y,Xiong G,et al. ULDC:Unsupervised Learning-Based Data Cleaning for Malicious Traffic With High Noise[J]. The Computer Journal,2024,67(3):976-987.

[14] Zhang C,Bengio S,Hardt M,et al. Understanding deep learning (still) requires rethinking generalization[J]. Communications of the ACM,2021,64(3):107-115.

[15] Fallah S,Bidgoly A J. Android malware detection using network traffic based on sequential deep learning models[J]. Software:Practice and Experience,2022,52(9):1987-2004.

[16] Han B,Yao J,Niu G,et al. Masking:A new perspective of noisy supervision[EB/OL]. http://arxiv.org/abs/1805.08193,2018-05-21/2023-06-01.

[17] Lyu Y,Tsang I W. Curriculum loss:Robust learning and generalization against label corruption[J]. arXiv preprint arXiv:1905.10045,2019:1-2.

[18] Wang Y,Sun X,Fu Y. Scalable penalized regression for noise detection in learning with noisy labels[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022:346-355.

[19] Weisberg S. Applied linear regression[M]. John Wiley & Sons,2005:1-3.

[20] Huber P J. Robust statistics[M]. John Wiley & Sons,2004:2-4.

[21] Wang W,Zhu M,Zeng X,et al. Malware traffic classification using convolutional neural network for representation learning[C]. 2017 International conference on information networking (ICOIN). IEEE,2017:712-717.

High Noise Traffic Classification based on Penalty Regression

BAI Kaiyi[1], SHENG Zhiwei[1,2], HUANG Yuanyuan[1,2]

(1. College of Cyberspace Security,Chengdu University of Information Technology,Chengdu 610225,China;2. Sichuan Provincial Key Laboratory of Advanced Cryptography Technology and System Security,Chengdu 610225,China)

Abstract:This paper aims to address the data quality issues faced in the field of high noise traffic classification. In response to the reality that network traffic data is prone to interference,the idea of noisy label learning (LNL) is introduced,and noise is artificially added to blur features. Firstly,establish a linear relationship between features and labels,and then use non-zero mean-shift parameters to identify noisy data. Simulate various interference information in real situations by manually adding symmetric and asymmetric noise. Therefore,this paper proposes a high-noise traffic classification model based on L2 regularization (PR-2),which converts traffic into images and applies the L2 regularization method to process noisy labels to improve the performance of the classification model under high-noise traffic. The effectiveness of this method was validated on the USTC-TF2016 dataset and compared with LSTM,BiTCN,BoAu,CL,INCV,and FINE methods. The experimental results show that the PR-2 method can still achieve 95.16% and 86.15% accuracy even when the proportion of symmetric and asymmetric noise is 80%,demonstrating its effectiveness and usability in processing high-noise data.

Keywords:high-noise traffic;traffic classification;deep learning;learning with noisy labels