Pure Mathematics 理论数学, 2024, 14(1), 23-33
Published Online January 2024 in Hans. https://www.hanspub.org/journal/pm
https://doi.org/10.12677/pm.2024.141004

Hans汉斯

加密流量数据集类别不平衡的研究

王　晓

上海理工大学光电信息与计算机工程学院，上海

收稿日期：2023年12月3日；录用日期：2024年1月4日；发布日期：2024年1月11日

摘　要

近年来，随着深度学习技术的迅猛发展，网络安全领域的研究人员开始探索利用深度学习解决加密流量分类问题。然而，目前公开的加密流量数据集存在严重的类别不平衡问题，这对于深度学习分类方法的性能造成了一定的影响。从头构建一个完整的加密流量数据集是耗时且昂贵的。为了克服这个问题，本文提出了一种基于改进的生成对抗网络(GAN)的加密流量生成模型。该模型通过在GAN模型中添加数据包的统计特征和网络流的类别标签作为条件约束，从而生成逼真的流量数据，进而扩充数据集。实验证明，在使用经过本文方法增强的数据集时，基于深度学习的加密流量分类器展现出比使用随机过采样(ROS)、合成少数类过采样技术(SMOTE)和传统的对抗生成网络(GAN)技术更出色的性能。

关键词

加密流量分类，平衡数据集，深度学习，生成对抗网络

Research on Class Imbalance in Encrypted Traffic Datasets

Xiao Wang

School of Optical-Electrical and Computer Engineering, University of Shanghai for Science and Technology, Shanghai

Received: Dec. 3rd, 2023; accepted: Jan. 4th, 2024; published: Jan. 11th, 2024

Abstract

In recent years, with the rapid development of deep learning technology, researchers in the field of network security have begun to explore using deep learning to solve the problem of encrypted traffic classification. However, currently available encrypted traffic datasets suffer from serious class imbalance issues, which can adversely affect the performance of deep learning classification methods. Creating a complete encrypted traffic dataset from scratch is both time-consuming and

文章引用: 王晓. 加密流量数据集类别不平衡的研究[J]. 理论数学, 2024, 14(1): 23-33.
DOI: 10.12677/pm.2024.141004

expensive. To address this issue, this paper proposes an improved generative adversarial network (GAN) based model for generating encrypted traffic data. The model adds packet statistics feature vectors as conditional constraints to the GAN model, thereby generating realistic traffic data to expand the dataset. Experimental results show that when using our method to enhance the dataset, the deep learning-based encrypted traffic classifier exhibits better performance than that using random oversampling (ROS), synthetic minority oversampling technique (SMOTE), and traditional GAN techniques.

Keywords

Encrypted Traffic Classification, Balanced Dataset, Deep Learning, Generative Adversarial Networks

1. 引言

随着深度学习技术在网络安全领域的广泛应用，特别是网络入侵检测与网络加密流量识别中，类别不平衡问题会对深度学习模型的性能产生消极影响，所以处理加密流量数据集类别不平衡的问题变得尤为重要[1]。

在解决类别不平衡问题中，采样是一种常用的方法之一，包括过采样和欠采样。过采样通过增加少数类别样本的复制数量来平衡数据集，而欠采样则是通过减少多数类别样本的数量来达到平衡[2]。如ROS (Random Over-Sampling)是一种过采样技术，它通过复制少数类别中的样本来增加其数量，以达到平衡数据集的目的。ROS 通过在少数类别中随机选择样本，并将其复制多次来进行重采样。这样可以增加少数类别的样本数量，从而改善数据集的平衡性[3]。虽然 ROS 简单易用，但可能会导致过度拟合的风险，因为重复采样样本可能使得模型过于关注少数类别的特征。修改损失函数也是一种常见的方法，通过给不同类别的样本赋予不同的权重或者引入正则化项来调整模型的训练过程，以便更加关注少数类别。但赋予不同类别的样本不同的权重可能会引入偏差。如果权重设置得不合理，可能会导致模型过度关注少数类别，而忽视多数类别。另一种解决类别不平衡问题的方法是生成合成数据。在这方面，SMOTE (Synthetic Minority Over-sampling Technique)是一种常用的技术[4]。SMOTE 通过随机选择少数类别中的样本，并根据其邻居样本的特征生成新的合成样本。具体来说，对于一个少数类别样本，SMOTE 会计算其与最近邻样本之间的差值，然后根据这个差值与一个随机数相乘的结果，生成一个新的合成样本。

近年来，基于生成对抗网络(GAN)的方法在生成合成数据方面取得了显著的进展，被广泛应用于解决类别不平衡问题[5]。GAN 通过训练一个生成器网络来生成具有少数类别特征的新样本，并与判别器网络进行对抗训练，从而提高模型在少数类别上的性能。一些研究人员为了克服网络数据不平衡问题，提出了基于 GAN 的方法来生成流量样本。ACGAN [6]用于生成合成的流量样本，以平衡知名的流量数据集 NIMS 中的次要和主要类别。具体而言，AC-GAN 同时采用随机噪声和类标签作为输入，以根据输入类别标签生成样本。实验结果表明，他们的方法相比于 SMOTE 等其他方法，具有更好的性能。然而，NIMS 数据集中仅包含 SSH 和非 SSH 两个类别。为此，基于上述研究的缺陷，本文设计了生成器和判别器，并在此基础上添加了条件约束，通过学习原始流量数据的特征来生成合成流量样本。然后，将合成数据与

原始(即真实)数据相结合，构建了一个新的流量数据集。这种方法旨在解决多类别不平衡问题，并且可应用于其他类型的数据集。

2. 本文方法

2.1. 数据处理

本文设计的数据处理如图 1 所示，旨在实现对真实网络中原始流量数据的分割和预处理，以便进行后续的分析和处理。在真实的网络系统中，流量数据源不包含单个应用程序的有序序列，而是由同一网络段上不同主机发送的各种数据包组成。因此，在这些原始流量中区分出由单个应用程序产生的流量就需要进行流量分割。每个流量都可以表示为具有相同的五元组(源 IP、源端口、目的 IP、目的端口、传输层协议)的数据包序列[7]。考虑到隐私问题，在处理原始流量之前，需要使用 TraceWrangler.exe 工具将 IP 地址匿名化，然后根据五元信息进行流量分割。

由于 DNS、TCP 三次握手等数据包与应用程序识别无关，将这些无关的数据包进行过滤从而消除数据集中的噪音。网络中的其他层主要包含网络控制和网络传输的信息而非有用的数据信息，因此只需关注应用层即数据包的有效载荷。注意到网络流中包含的分组数量和每个分组的大小通常是不确定的，一种常见的方法是统一数据包长度以及数据包个数选择网络流的前 n 个数据包，并保留每个分组的前 m 个字节。如果网络流中的数据包数量超过了 n，需要对其进行截断；反之，如果数据包数量不足 n，则需用 0 进行填充。同样地，如果一个数据包中的字节数多于或少于 m，也需进行相同的处理预处理，以提供准确而一致的数据。最终，通过网络流的矢量化，将字节单位转换为 0 到 255 的整数，并进行归一化操作来得到最终的网络流量特征图。

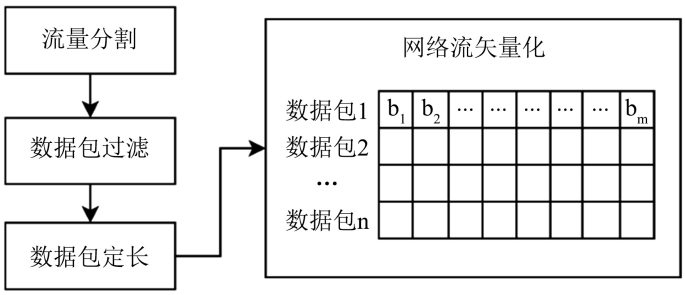

Figure 1. Data processing flow
图 1. 数据处理流程

2.2. 条件约束 GAN 网络

条件约束 GAN (Conditional GAN，简称 CGAN)是一种基于生成对抗网络(GAN)的深度学习模型，它在原始的 GAN 模型基础上增加了条件约束，在生成器和判别器之间引入了额外的网络流的统计特征，通过该附加条件来控制生成器产生样本，从而实现更加精细化的生成任务。CGAN 最初由 Mirza 和 Osindero 在 2014 年提出[8]，其主要思想是将条件信息 c 作为噪声 z 和生成器 G 的输入进行联合训练，同时将条件信息 c 与真实样本 x 和判别器 D 的输入一同输入。

本文的模型框架如图 2 所示，首先，定义生成器 G 和判别器 D。G 接受一个噪声向量 z 作为输入，生成伪造的网络流量样本 x_{fake}，即 $x_{fake}=G(z)$。D 接受一个网络流量样本 x 作为输入，输出一个标量值 $D(x)$ 表示该样本是真实网络流量的概率。GAN 的目标是最小化生成器和判别器之间的损失函数 $V(G,D)$。其中，判别器的目标是使得真实网络流量的 $D(x)$ 尽可能接近于 1，而生成器的目标是使得它生成的网络流

量的 $D\left(x_{fake}\right)$ 尽可能接近于 1。因此，可以将 $V\left(G,D\right)$ 表示为以下公式：

$$V\left(G,D\right)=\min_{G}\max_{D}\mathbb{E}_{x\sim p_{data}(x)}\left[\log D\left(x\right)\right]+\mathbb{E}_{z\sim p_{z}(z)}\left[\log\left(1-D\left(G\left(z\right)\right)\right)\right]$$

其中 $\mathbb{E}$ 表示期望值，$p_{data}\left(x\right)$ 表示真实网络流量的分布，$p_{z}\left(z\right)$ 表示噪声向量 z 的分布。$\log D\left(x\right)$ 表示 $D\left(x\right)$ 的对数，$\log\left(1-D\left(G\left(z\right)\right)\right)$ 表示 $D\left(G\left(z\right)\right)$ 的对数差。

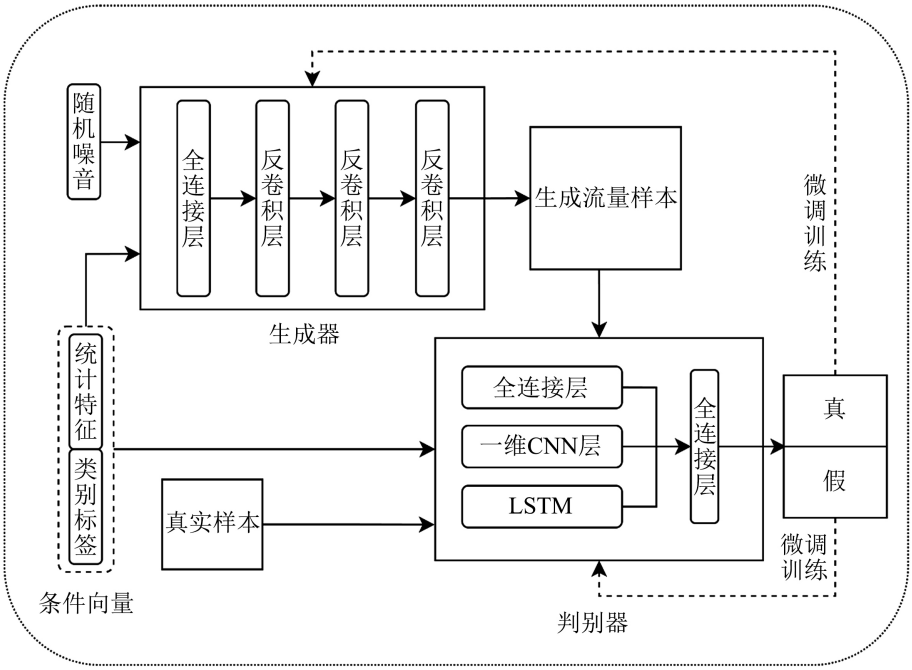

Figure 2. The overall framework of this method
图 2. 本文方法整体框架

为了使生成的网络流量更加逼真，可以在训练过程中引入条件约束。具体地，将真实网络流量的统计特征和类别标签作为额外的条件输入提供给生成器和判别器。假设条件向量用向量 c 来表示，则可以将生成器的输入修改为一个元组 $\left(c,z\right)$，生成器将噪声向量 z 和条件向量 c 映射为虚假的网络流量样本 x_{fake}。判别器的输入也需要加入条件向量 c，即 $D\left(x,c\right)$ 表示网络流量样本 x 和条件向量 c 的区分能力。此时，GAN 的目标函数变为：

$$V\left(G,D\right)=\min_{G}\max_{D}\mathbb{E}_{(x,c)\sim p_{data}(x,c)}\left[\log D\left(x,c\right)\right]+\mathbb{E}_{z\sim p_{z}(z)}\left[\log\left(1-D\left(G\left(z,c\right)\right)\right)\right]$$

其中 $\mathbb{E}$ 表示期望值，$p_{data}\left(x,c\right)$ 表示真实网络流量和条件的联合分布，$p_{z}\left(z\right)$ 表示噪声向量 z 的分布。$\log D\left(x,c\right)$ 表示 $D\left(x,c\right)$ 的对数，$\log\left(1-D\left(G\left(z,c\right)\right)\right)$ 表示 $D\left(G\left(z,c\right)\right)$ 的对数差。

通过不断迭代训练生成器和判别器，可以逐渐提高生成的网络流量样本的质量，使其更加逼真。

2.3. 条件向量

采用流量的统计特征和类别标签作为条件向量，可以提供全面且综合的信息来描述网络流量。统计特征可以反映流量的基本属性和行为特征，而类别标签可以表示流量所属的具体类别。条件向量作为生成对抗网络的输入，可以帮助网络更好地学习和理解网络流量的特征和类别。

在本研究中选择了表 1 所示的一组典型的统计特征作为条件向量的一部分。假设流量的统计特征向量

f，类别标签通常使用独热编码向量 I 来表示。独热编码是一种将离散值表示为二进制向量的方法，其中只有一个元素为 1，其余元素为 0。每个类别对应一个独热编码向量，向量的长度等于类别的总数。

$$f=[f_1,f_2,\cdots,f_n]$$

$$I_i=[0,0,\cdots,1,\cdots,0]$$

其中，f_i 表示第 i 个统计特征归一化后的值，I_i 的第 i 个元素为 1，表示该样本属于第 i 个类别，其余元素都为 0。

生成对抗网络的输入条件向量 c 可以通过将流量的统计特征和类别标签进行拼接得到，即 $c=[f,I]$，这样的拼接操作可以保留并利用统计特征与类别标签之间的信息关联。

Table 1. Statistical characteristics
表 1. 统计特征

特征	说明	特征	说明
f_1	流持续时间	f_{23}	双向流数据包的大小标准差
f_2	源到目标的数据包数量	f_{24}	双向流最小数据包到达时间间隔
f_3	目标到源的数据包数量	f_{25}	双向流平均数据包到达时间间隔
f_4	源到目标的最大数据包大小	f_{26}	双向流数据包到达时间间隔标准差
f_5	源到目标的最小数据包大小	f_{27}	双向流最大数据包到达时间间隔
f_6	源到目标的平均数据包大小	f_{28}	源到目的流最小数据包到达时间间隔
f_7	源到目标的数据包大小标准差	f_{29}	源到目的流平均数据包到达时间间隔
f_8	目标到源的最大数据包大小	f_{30}	源到目的流数据包到达时间间隔标准差
f_9	目标到源的最小数据包大小	f_{31}	源到目的流最大数据包到达时间间隔
f_{11}	目标到源的平均数据包大小	f_{32}	目的到源流最小数据包到达时间间隔
f_{12}	目标到源的最小数据包大小	f_{33}	目的到源流平均数据包到达时间间隔
f_{13}	目标到源的数据包大小标准差	f_{34}	目的到源流数据包到达时间间隔标准差
f_{14}	双向流最小数据包大小	f_{35}	目的到源流最大数据包到达时间间隔
f_{15}	双向流平均数据包大小	f_{36}	每秒传输的数据包字字节大小
f_{16}	双向流最大数据包大小	f_{37}	每秒传输的数据包数
f_{17}	URG 标志位的数据包数量	f_{38}	源端口号
f_{18}	ACK 标志位的数据包数量	f_{39}	目的端口号
f_{19}	PSH 标志位的数据包数量	f_{40}	协议号
f_{20}	RST 标志位的数据包数量	f_{41}	SYN 标志位的数据包数量
f_{21}	FIN 标志位的数据包数量	f_{42}	CWR 标志位的数据包数量
f_{22}	双向流 TCP 标志位计数器	f_{43}	ECE 标志位的数据包数量

2.4. 生成器网络结构

生成器的作用是将给定的条件向量 c 和噪声向量 z 转换为逼真的加密流量样本 x。生成器的设计采用了多层反卷积网络[9]，通过逐步上采样来生成与原始加密流量样本相似的时间序列数据。将生成器表示为一个非线性映射函数 G，如下所示：

$$x = G(c, z)$$

首先，将条件向量 c 与噪声向量 z 进行拼接，得到一个低维的中间特征向量 h_0。这一步使用一个全连接层实现。

$$h_0 = \mathrm{ReLU}\left(W_{cz}[c, z] + b_{cz}\right)$$

其中，$[c, z]$ 表示将条件向量 c 和噪声向量 z 进行拼接，W_{cz} 和 b_{cz} 分别表示全连接层的权重和偏置。

接下来，通过多个反卷积层进行逐步上采样，将中间特征向量 h_0 转换为最终的生成样本 x。

$$h_1 = \mathrm{Deconv}(h_0, W_1, b_1)$$

$$h_2 = \mathrm{Deconv}(h_1, W_2, b_2)$$

$$\vdots$$

$$h_n = \mathrm{Deconv}(h_{n-1}, W_n, b_n)$$

$$x = \mathrm{Tanh}(h_n)$$

其中，Deconv 表示反卷积层的操作，$W_1, W_2, \cdots, W_n$ 和 $b_1, b_2, \cdots, b_n$ 分别表示反卷积层的权重和偏置。每个反卷积层都会增加一倍的特征图尺寸，以逼近原始加密流量样本的维度。最后一层使用 tanh 作为激活函数，将生成样本 x 的取值限制在−1 到 1 之间。

2.5. 判别器网络结构

判别器的作用是将生成的加密流量样本 x 与真实的加密流量样本进行区分。判别器包含三个子网络，用于提取加密流量样本 x 和条件向量 c 的高层特征：全连接层、一维卷积层[10]和 LSTM 层[11]。将判别器表示为一个非线性映射函数 D，如下所示：

$$y = D(x, c)$$

具体而言，使用全连接层提取条件向量 c 的高层特征，使用一维卷积层提取加密流量样本 x 的空间特征，使用 LSTM 层提取时间特征。最后，将三部分特征进行融合，并使用全连接层进行二元分类，判断输入数据是真实数据还是生成数据。

首先，使用全连接层提取条件向量 c 的高层特征，得到向量 h_c。

$$h_c = \mathrm{ReLU}\left(W_c[c] + b_c\right)$$

接下来，使用一维卷积层提取加密流量样本 x 的空间特征，得到向量 h_x

$$h_x = \mathrm{Conv1D}(x)$$

然后，使用 LSTM 层提取加密流量样本 x 的时间特征，得到向量 h_t

$$h_t = \mathrm{LSTM}(x)$$

最后，将三部分特征进行拼接，得到向量 h。

$$h = \text{Concat}\left(h_c, h_x, h_t\right)$$

通过一个全连接层和 sigmoid 激活函数，将向量 h 映射为一个二元分类结果 y，表示输入数据是真实数据还是生成数据。

$$y = \text{Sigmoid}\left(W_y\left[h\right] + b_y\right)$$

其中 W_c, W_x, W_t, W_y 和 b_c, b_y 分别表示全连接层的权重和偏置。整个判别器网络的目标是最大化生成数据与真实数据之间的差异，从而使生成器能够生成更逼真的加密流量样本。

2.6. 模型训练

在训练过程中，采用交替优化的策略。首先，固定生成器参数，通过最小化判别器来更新判别器参数，使判别器能够更好地区分真实样本和伪造样本。然后，固定判别器参数，通过最小化生成器损失函数来更新生成器参数，使生成器能够生成更逼真的伪造样本。这样反复迭代优化，直到生成器和判别器达到一定的平衡状态。具体地，将训练过程分为两个阶段：预训练和对抗训练。在预训练阶段，通过最小化生成器和判别器之间的均方误差来训练模型。在对抗训练阶段，使用 GAN 的目标函数进行训练。

首先，随机初始化生成器 G 和判别器 D 的参数。然后，使用真实的加密流量样本 x 和条件向量 c 作为输入，将其通过生成器 G 生成虚假的加密流量样本 x_{fake}。使用均方误差(MSE)损失函数来最小化生成器和判别器之间的误差：

$$\mathcal{L}_{pre} = \frac{1}{N}\sum_{i=1}^{N}\left\|D\left(x_i, c_i\right) - 1\right\|^2 + \frac{1}{N}\sum_{i=1}^{N}\left\|D\left(G\left(z_i, c_i\right), c_i\right)\right\|^2$$

其中，N 是训练样本数量，z_i 是从噪声分布 $p_z(z)$ 中采样得到的噪声向量。通过预训练，可以加快模型的收敛速度。

在对抗训练中，使用 GAN 的目标函数进行训练。具体来说，使用交替优化的方式来更新生成器和判别器的参数。在每次迭代中，首先固定生成器 G，最大化判别器 D 的目标函数，即：

$$\mathcal{L}_D = -\frac{1}{N}\sum_{i=1}^{N}\left[\log D\left(x_i, c_i\right) + \log\left(1 - D\left(G\left(z_i, c_i\right), c_i\right)\right)\right]$$

然后，固定判别器 D，最小化生成器 G 的目标函数，即：

$$\mathcal{L}_G = -\frac{1}{N}\sum_{i=1}^{N}\log D\left(G\left(z_i, c_i\right), c_i\right)$$

最终的损失函数为：

$$\mathcal{L}_{GAN} = \mathcal{L}_D + \mathcal{L}_G$$

在每次迭代中，从真实的加密流量样本和噪声分布中分别采样得到训练样本，并根据目标函数进行优化。

3. 实验

3.1. 实验配置

本实验的硬件环境为 NVIDIA GeForce RTX 3060 Laptop GPU，CUDA 版本为 11.6，每条网络流选择前 36 个数据包，每条数据包选择前 256 个字节。优化器采用 Adam 优化器，生成器的学习率为 0.0001，判别器的学习率为 0.0002，批次大小为 32，迭代次数为 3000 轮次。

3.2. 数据集

CIC-Darknet2020 数据集[12]是一个用于研究和分析 Darknet 流量的数据集，Darknet 是互联网未使用的地址空间，通常不与世界上其他计算机进行交互。该数据集通过合并 ISCXTor2016 [13]和 ISCXVPN2016 [14]两个公共数据集中的 Tor 和 VPN 流量，构建了一个包含了各种类型 Darknet 流量的完整数据集。在 CIC-Darknet2020 数据集中，Darknet 流量的详细信息如表 2 所示，该数据集的类别存在数据集不平衡的现象。

Table 2. Details of Darknet traffic in CIC-Darknet2020 data set
表 2. CIC-Darknet2020 数据集中 Darknet 流量的详细信息

类别	流数量	占比
Audio-Stream	13,284	54.64%
Browsing	263	1.08%
Chat	4541	18.68%
Email	582	2.39%
P2P	220	0.90%
File-Transfer	2610	10.74%
Video-Stream	1346	5.54%
VOIP	1465	6.03%

对 CIC-Darknet2020 数据集进行数据平衡，采用 ROS、SMOTE、GAN 和本文方法共四种生成方法使得每个类别的流量都有 3096 个样本，每类占比均为 12.5%。

3.3. 评估指标

在本研究中，采用了准确率和 F1 值作为分类模型的评估指标。其中准确率(Accuracy)是指分类器正确分类的样本数占总样本数的比例，其公式如下所示：

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

其中，TP 表示真正例(True Positive)，即实际类别为正例且预测为正例的样本数；TN 表示真负例(True Negative)，即实际类别为负例且预测为负例的样本数；FP 表示假正例(False Positive)，即实际类别为负例但预测为正例的样本数；FN 表示假负例(False Negative)，即实际类别为正例但预测为负例的样本数。

F1 值是综合考虑了分类器的精准率(Precision)和召回率(Recall)的一种度量指标。它能够同时兼顾分类器对于正例和负例的分类效果，其公式如下所示：

$$F1 = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}$$

其中，精准率(Precision)表示预测为正例的样本中实际为正例的比例，其公式如下所示：

$$\text{Precision} = \frac{TP}{TP + FP}$$

召回率(Recall)表示实际为正例的样本中被分类器正确预测为正例的比例，其公式如下所示：

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

需要注意的是，对于多分类模型而言，采用宏观平均(macro-averaging)来计算准确率和 F1 值。宏观平均先分别计算各个类别的准确率和 F1 值，再求它们的平均值。

3.4. 实验结果分析

基于 DFR [15]种使用 LSTM 的流量分类方法，表 3 展示在 CIC-Darknet2020 数据集上运用不同的类不平衡分类算法所得到的准确率。这些算法包括未经处理的不平衡数据集、基于 ROS 方法、基于 SMOTE 方法、基于 GAN 的方法以及本文所提出的方法。通过对比这些分类结果，可以看出本文提出的方法在该数据集上取得了明显优势，相较于不平衡数据集和其他三种过采样方法，其分类准确率更加突出，比未处理的数据集准确率提高了 7%。另外，在 CIC-Darknet2020 数据集种各类别的 F1 值如图 3 所示，可以看出本文方法可以显著提高少数类的 F1 值，在 Browsing、Email、P2P 这三种占比极少的类别中，F1 值比次优的 GAN 方法提升了 10%、2.10%、4.40%。

值得注意的是，传统的不平衡数据集处理方法存在一些局限性。例如，基于 ROS (Random Over Sampling)方法会简单地复制少数类样本以平衡数据集，但这容易导致过拟合或者丢失原始信息。而基于 SMOTE (Synthetic Minority Over-sampling Technique)方法虽然可以生成合成的少数类样本，但对于高维特征空间的数据集来说，其合成样本的质量可能无法保证。

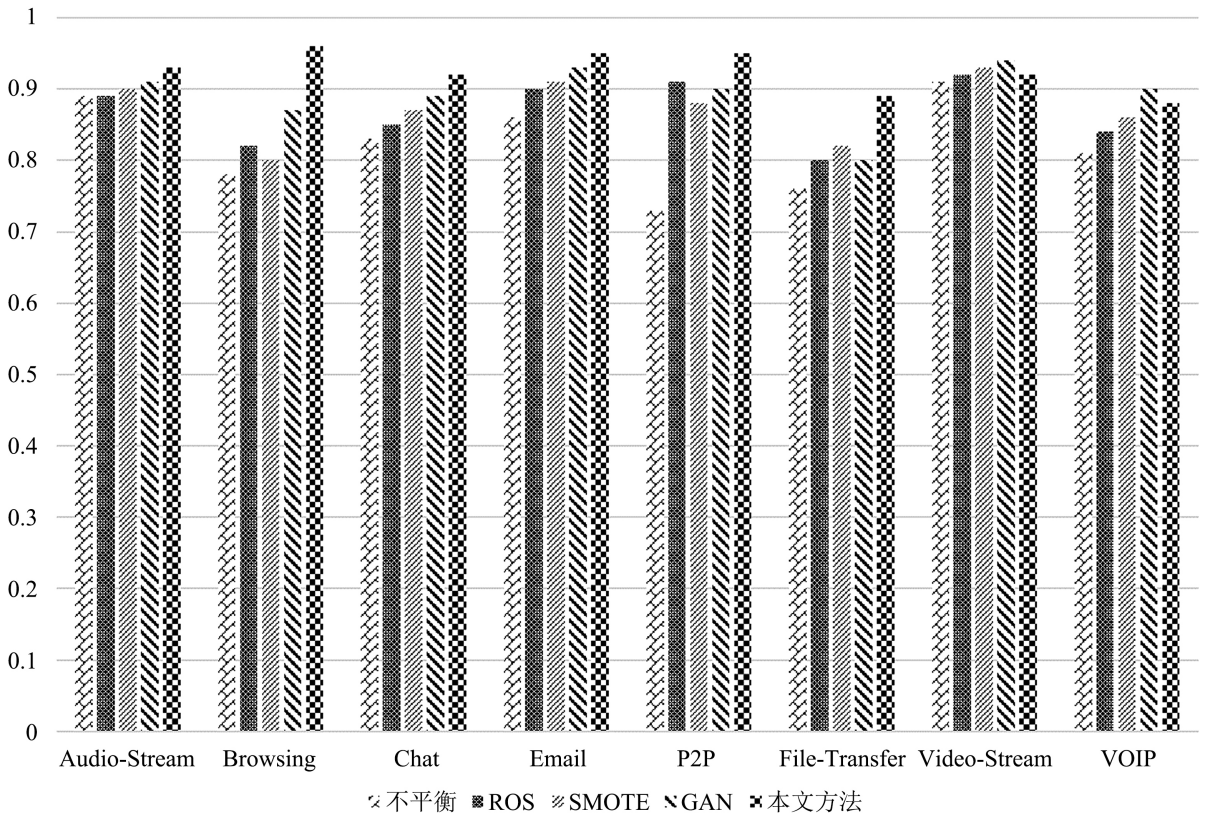

Figure 3. Comparison of F1 values of each category on the CIC-Darknet2020 data set
图 3. CIC-Darknet2020 数据集上各类别 F1 值对比

Table 3. Classification results on the CIC-Darknet2020 data set
表 3. 在 CIC-Darknet2020 数据集上的分类结果

方法	准确率
不平衡数据集	89.24%
ROS	92.19%
SMOTE	92.78%
GAN	93.64%
本文方法	95.49%

相比之下，本文通过引入额外的条件信息，使得生成网络更加可控，能够实现有监督学习的效果，并且在一定程度上缓解了 GAN 存在的训练不稳定、模式崩溃等问题。可以对生成过程进行更精细控制，充分利用统计特征和类别标签作为条件向量的 GAN 网络生成网络流量，能够更加有效地生成符合原始数据分布特征的合成样本，从而改善了数据集的平衡性并提升了分类效果。

4. 结论

通过本文对加密流量数据集类别不平衡问题的研究，我们提出了一种基于改进的生成对抗网络(GAN)的加密流量生成模型，旨在解决目前公开的加密流量数据集存在的严重类别不平衡问题。通过实验验证，该方法相对于传统的方法在分类准确率和 F1 值方面具有显著优势。通过引入额外的条件信息和结合统计特征和类别标签，使得生成网络变得更加可控，可以更准确地生成符合条件的合成网络流量数据，从而提高训练和生成效果，使生成的网络流量更加真实和逼真，从而改善了数据集的平衡性并提升了分类效果。这为解决实际网络流量分类中的类别不平衡问题提供了新的思路和方法。我们希望未来能够进一步探索和完善这一方法，为网络安全领域提供更加有效的解决方案。

参考文献

[1] Vu, L., Van Tra, D. and Nguyen, Q.U. (2016) Learning from Imbalanced Data for Encrypted Traffic Identification Problem. *Proceedings of the Seventh Symposium on Information and Communication Technology*, *ser*. *SoICT*'16, New York, NY, 147-152. https://doi.org/10.1145/3011077.3011132

[2] Japkowicz, N. (2000) Learning from Imbalanced Data Sets: A Comparison of Various Strategies. *AAAI Workshop on Learning from Imbalanced Data Sets*.

[3] Chawla, N.V., Bowyer, K.W., Hall, L.O., *et al.* (2002) SMOTE: Synthetic Minority Over-Sampling Technique. *Journal of Artificial Intelligence Research*, **16**, 321-357. https://doi.org/10.1613/jair.953

[4] Wang, Q., Li, L., Jiang, B., *et al.* (2020) Malicious Domain Detection Based on K-Means and Smote. *International Conference on Computational Science*, Amsterdam, The Netherlands, Springer, Cham, 468-481. https://doi.org/10.1007/978-3-030-50417-5_35

[5] Goodfellow, I., Pouget-Abadie, J., Mirza, M., *et al.* (2014) Generative Adversarial Nets. *Advances in Neural Information Processing Systems*, Montreal, 2672-2680.

[6] Vu, L., Bui, C.T. and Nguyen, Q.U. (2017) A Deep Learning Based Method for Handling Imbalanced Problem in Network Traffic Classification. *Eighth International Symposium on Information & Communication Technology*, New York, December 2017, 333-339. https://doi.org/10.1145/3155133.3155175

[7] Dainotti, A., Pescape, A. and Claffy, K.C. (2012) Issues and Future Directions in Traffic Classification. *Network IEEE*, **26**, 35-40. https://doi.org/10.1109/MNET.2012.6135854

[8] Mirza, M. and Osindero, S. (2014) Conditional Generative Adversarial Nets. *Computer Science*, 2672-2680.

[9] Zeiler, M.D., Krishnan, D., Taylor, G.W., *et al.* (2010) Deconvolutional Networks. *Computer Vision & Pattern Recognition*, San Francisco, CA, 13-18 June 2010, 2528-2535. https://doi.org/10.1109/CVPR.2010.5539957

[10] Wang, W., Zhu, M., Wang, J., *et al.* (2017) End-to-End Encrypted Traffic Classification with One-Dimensional Convolution Neural Networks. *IEEE International Conference on Intelligence and Security Informatics* (*ISI*), Beijing, 22-24 July 2017, 43-48. https://doi.org/10.1109/ISI.2017.8004872

[11] Lin, K., Xu, X. and Gao, H. (2021) TSCRNN: A Novel Classification Scheme of Encrypted Traffic Based on Flow Spatiotemporal Features for Efficient Management of IIoT. *Computer Networks*, **190**, Article ID: 107974. https://doi.org/10.1016/j.comnet.2021.107974

[12] Lashkari, A.H., Kaur, G. and Rahali, A. (2020) DIDarknet: A Contemporary Approach to Detect and Characterize the Darknet Traffic Using Deep Image Learning. *Proceedings of the* 2020 10*th International Conference on Communication and Network Security* (*ICCNS* 2020), New York, 27-29 November 2020, 1-13.

[13] Lashkari, A.H., Draper-Gil, G., Mamun, M.S.I., *et al.* (2016) Characterization of Encrypted and VPN Traffic Using Time-Related Features. *Proceedings of the* 2*nd International Conference on Information Systems Security and Privacy ICISSP*, **1**, 407-414. https://doi.org/10.5220/0005740704070414

[14] Lashkari, A.H., Gil, G.D., Mamun, M.S.I., *et al.* (2017) Characterization of Tor Traffic Using Time Based Features. *International Conference on Information Systems Security & Privacy*, Porto, 253-262.

[15] Zeng, Y., Gu, H., Wei, W., *et al.* (2019) Deep-Full-Range: A Deep Learning Based Network Encrypted Traffic Classification and Intrusion Detection Framework. *IEEE Access*, **7**, 45182-45190. https://doi.org/10.1109/ACCESS.2019.2908225