文章编号：2095-2163(2024)03-0181-06 中图分类号：TP391 文献标志码：A

基于置信度与级联结构的未知网络流量检测

吴志远，董育宁，李 涛
（南京邮电大学 通信与信息工程学院，南京 210003）

摘 要：为了提升开集流识别性能，本文在对已知类和新类的置信度分布分析基础上，提出一种基于置信度信息与级联结构的未知网络流量检测方法。该方法通过级联结构，先将具有高置信度的新类样本检测出来；利用最大置信度差对新类和已知类进行分类；利用最大置信度对已知类进行细分类。为了更好地检测高置信度新类，还设计了从未标记数据筛选伪负样本的算法。实验表明，与现有代表性方法相比，本文方法的已知类 $F1$ 提高约 13%，新类 $F1$ 提高约 3%，总体准确率提高约 5%，训练和分类耗时也明显少于现有方法。

关键词：开集流识别；置信度；未知网络流量检测；未标记数据

Unknown network traffic detection based on confidence information and cascade structure

WU Zhiyuan, DONG Yuning, LI Tao
(School of Communications and Information Engineering, Nanjing University of Posts and Telecommunications, Nanjing 210003, China)

Abstract: In order to improve the performance of open set flow recognition, this paper proposes an unknown network traffic detection method based on confidence and cascade structure, based on the analysis of confidence distribution of known and new classes. This method uses a cascade structure to firstly detect new class samples with high confidence, then uses the maximum confidence difference to classify the new and known classes, and uses the maximum confidence to finely classify the known classes. In order to better detect new classes with high confidence, an algorithm for filtering pseudo negative samples from unlabeled data is also designed. The experiment shows that compared with the existing representative method, the $F1$ of known class is increased by 13%, and the $F1$ of new class is increased by 3%, and the overall accuracy is increased by 5%. Training and classification are also significantly less time-consuming than existing method.

Key words: open set flow recognition; confidence; unknown network traffic detection; unlabeled data

0 引 言

网络流量分类对网络管理非常重要，如服务质量保证、网络资源分配等[1]。随着互联网和多媒体技术的快速发展，网络流量类别不断增多，新型网络应用层出不穷。对于已训练好的分类器来说，这些新型应用流量是未知的，会被错误地分类，不能识别出新类样本。机器学习和深度学习方法一般需要大量标记样本作为支撑，但在真实网络场景下无法获得未知类的标记样本。对于包含未知类的流识别问题被称为开集流识别(Open Set Flow Recognition, OSFR)[2]。

对于开集流识别和分类问题，目前已有的方法包括利用单类样本分类(不需要负样本)来处理新类检测，此类方法虽然可以检测新类，但将已知类都归为一类[3-4]；另一种途径是通过对抗学习以无监督/半监督方式生成负样本，但该方法虽然可以较好地检测新类样本，但容易将较多已知类样本错分到新类中，并且负样本生成时间一般较长[5-6]。

相对而言，基于对抗生成样本的方法在新类检测总体效果较好。受启发于对抗生成负样本的思路，同时针对这类方法现存的不足，本文设再叠加一级识别器，以过滤掉容易与已知类混淆的新类样本，但对抗生成样本的时间较长，而现实世界中较易获得未标记数据样本，可以从中选择伪负样本来辅助识别器的训练，达到新类检测效果好和分类时间短的双赢目的。

作者简介：吴志远(1998-)，男，硕士研究生，主要研究方向：网络流分类、多媒体通信与无线网络；李 涛(1964-)，男，副教授，硕士生导师，主要研究方向：新型网络及其路由技术。
通讯作者：董育宁(1955-)，男，博士，教授，博士生导师，主要研究方向：网络流分类、多媒体通信与无线网络。Email：19900011@ njupt.edu.cn

收稿日期：2023-03-17

1 相关工作

开集流识别是从真实网络角度提出的研究方向,要求分类模型在没有任何辅助信息时不仅能对已知类进行分类,还能准确识别新类[7]。近年来,开集流识别问题受到越来越多学者的关注,是当前机器学习领域中的热点。基于传统机器学习的开集流识别算法主要以基于支持向量机的方法为代表,此外,对抗学习思想也在解决开集流识别问题中得到广泛应用。

1.1 基于 SVM 的开集流识别

Scheirer 等学者[8]提出的 1-vs-Set 机制基于支持向量机(Support Vector Machine,SVM)对已知类信息占据的空间进行约束,以减小开放空间风险,应对开放环境中的单类识别问题。为了解决开放环境中的多类识别问题,Scheirer 等学者[9]利用紧凑衰减概率模型和极值理论进行概率估计,提出韦伯校准 SVM(Weibull-calibrated SVM, W-SVM)。PI-SVM[3]以多分类 SVM 为基础,并运用极值理论对决策边界上的正训练样本建模,克服了多类 W-SVM 中开放性对于阈值选取的影响。Scherreik 等学者[10]提出的概率开集 SVM(Probabilistic Open Set-SVM, POS-SVM),进一步研究了 W-SVM 和 PI-SVM 阈值设置问题,为每个已知类确定唯一的拒绝阈值。

1.2 基于对抗学习的开集流识别

Neal 等学者[11]借助生成对抗网络对训练集样本进行扩充,用生成对抗网络产生伪开放集样本,这些样本很接近但不属于已知类。Yang 等学者[6]提出的框架不仅可以生成负样本,还能在已知类样本较少时,产生已知类的正样本。Yang 等学者[12]基于生成对抗网络,生成与目标样本高度相似的样本作为负样本,还重新设计了判别器以区分已知类和新类,并对已知类细分类。

1.3 基于集成思想的开集流识别

Vareto 等学者[13]从已知类和新类来自不同的数据分布入手,着眼于相关的特征差异性,在识别上结合哈希函数和分类方法,并通过实验展示了探测样本时响应值直方图对两者的不同表现。Neira 等学者[14]将不同的分类器和特征结合起来,将一个新提出的基于开放集图的最优路径树分类器、遗传算法和多数投票技术融合在一起。Dong 等学者[15]提出一种将测试样本划分为已知、未知和不确定域的域划分算法,引入 bootstrapping 和 K-S Test,用于挖掘和微调每个域的决策边界。

尽管上述方法在应对开集流识别问题上取得了一定成效,但是也存在一些不足。基于支持向量机的开集流识别,虽然结构简单,但分类精度有待提高;基于集成思想的开集流识别虽然整体分类精度较高,但集成方法较为复杂。为此,本文提出一种基于置信度信息与级联结构的未知网络流量检测方法,该方法利用伪负样本训练的二分类器和最大置信度差来检测新类,再利用最大置信度对已知类进行细分类。

2 本文方法

2.1 总体框架

本文从已知类和新类的置信度分布差异入手,利用最大置信度差($CfD_{\max}$)区分已知类和新类,通过对两者的置信度分布差异深入分析得到 $CfD_{\max}$ 阈值 β。当样本的 $CfD_{\max}$ 大于 β 时,该样本被判为已知类;否则,判为新类。对于已知类样本,再根据其最大置信度($Cf_{\max}$)进行细分类。

由于部分新类样本的 $CfD_{\max}$ 也会大于阈值,本文利用分类器本身的特性进行区分。为了定位无法用阈值检测的新类,叠加了一级识别器,用已知类样本和筛选后的未标记流样本进行训练,训练阶段示意如图 1 所示。图 1 中,H_1 为随机森林(Random Forest, RF)二分类器,用来检测 $CfD_{\max}$ 大于阈值的新类;RF 多分类器 H_2 由已知类细分样本进行训练;在测试阶段,利用 β 进一步区分已知类和新类。

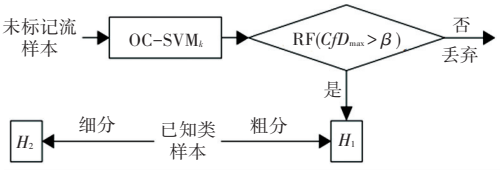

图 1 训练阶段

Fig. 1 Training stage

测试阶段示意如图 2 所示。图 2 中,x 是输入流数据,需要对其进行特征提取(Feature Extraction, FE)和特征选择(Feature Selection, FS);z 是输出的已知类样本标签,y_1 和 y_2 分别是检测出的不满足和满足阈值条件的新类样本。

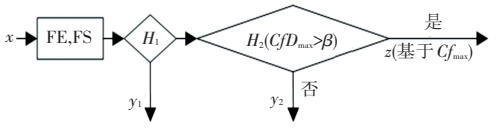

图 2 测试阶段

Fig. 2 Test stage

2.2　数据集

实验中采用 2 个真实网络数据集：ISCX non-VPN (ISCX)数据集[16] 和 VideD 视频数据集，其中 VideD 是 2021~2022 年期间在南京邮电大学校园网使用 Wireshark 软件采集得到，数据集的具体信息见表 1 和表 2。为了验证本文方法的普适性，将 ISCX 和 VideD 数据集合并组成混合数据集(MixD)，进行实验验证。

表 1　ISCX 部分数据

Table 1　Partial data of ISCX

类别	应用	样本数
文件传输	BitTorrent、Ftp、Skype	1 000
语音通话	Facebook、Hangouts、Skype	1 000
邮件	Email	1 000
聊天	Facebook	1 000
视频	Skype、Youtube	1 000

表 2　VideD 数据集

Table 2　VideD dataset

类别	应用	样本数
直播	douyu_480p、huya_480p、douyu_1080p	1 500
点播	tencent_480p、tencent_720p、youku_720p	1 500

2.3　特征提取

为了实现快速在线流分类，将流数据划分为1 s 的流段，再利用流段的前 10 个数据包进行特征计算，基于前 10 个数据包的计算特征进行分类。

采集的数据包括包大小序列、包到达时间序列、时间戳序列、包差值序列、上行速率系列和下行速率序列。分别对这 6 个序列计算的 17 个统计特征见表 3。此外，根据文献[17]提出的条件频率特征、即相邻到达的 2 个数据包的包大小等级条件频度，计算出 25 个下行条件频率特征和 4 个上行条件频率特征，共得到 131 个流特征。

表 3　流(包序列)统计特征

Table 3　Statistical characteristics of stream (packet sequence)

序号	特征名称	序号	特征名称
1	均值	6	最小值
2	标准差	7	奇异值个数
3	峰度	8	众数占比
4	偏度	9~17	10~90 分数
5	最大值		

2.4　特征选择

在线分类要求特征提取尽可能快，故进行特征选择和降维。步骤如下：

(1)进行特征提取时间复杂度分析，选出复杂度不超过 $O(n)$ (n 为数据包个数) 的特征子集；

(2)计算每个特征关于标签以及特征之间的皮尔森相关系数，对于皮尔森相关系数大于 0.9 的特征对，从中删除与标签相关性较小的一个；

(3)通过随机森林对剩下的特征进行排序，再根据重要性程度逐个增加特征，并观察分类准确率的变化，寻找性能的拐点，获得最优特征子集。

2.5　置信度信息分布

基于已知类训练集得到的模型，对已知类和新类输入数据会呈现不同的输出映射。本文利用随机森林对不同映射表现出的分类置信度分布进行分析，找出规律；从混合数据集(MixD)中随机选取 8 个类别样本作为已知类和另外 8 个作为新类，其置信度分布以及最大置信度差分布，分别如图 3 和图 4 所示。

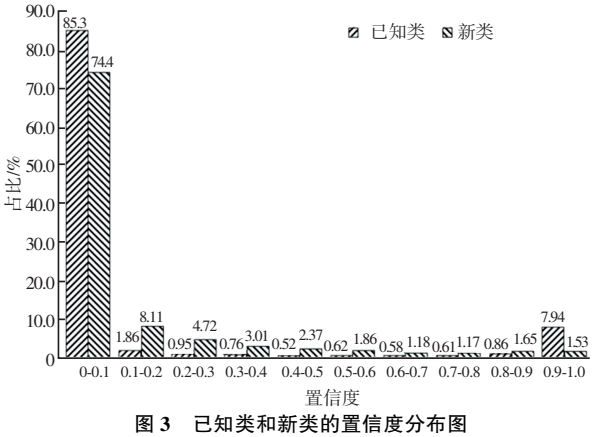

图 3　已知类和新类的置信度分布图

Fig. 3　Confidence distribution of known and new classes

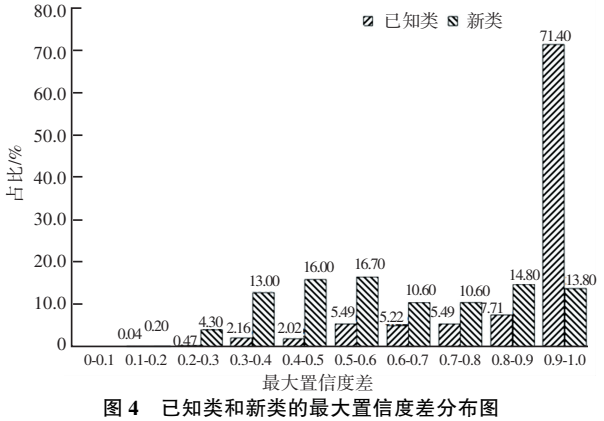

图 4　已知类和新类的最大置信度差分布图

Fig. 4　Maximum confidence difference distribution of known and new classes

由图 3 可知，已知类在最低置信度区间[0, 0.1]和最高置信度区间(0.9, 1.0]的占比均高于新类；由图 4 可知，已知类的 $CfD_{\max}$ 在(0.9, 1.0]之间的占比远远高于新类。故可利用 $CfD_{\max}$ 对已知类和新类进行区分，并选取阈值为 0.9。

为了验证最大置信度差阈值选取的普适性，本文还从 MixD 中随机选取若干不同已知类和新类组合进行广泛实验，结果见表 4。

从表 4 可知，对于不同类别的已知类和新类组合，已知类的最大置信度差在(0.9, 1.0]之间的占

比均高于 50%,而新类则均低于 50%,且已知类在此最大置信度差区间的占比远高于新类,故阈值 0.9 基本可有效区分已知类和新类。

表 4 混合数据集中不同已知类和新类组合的最大置信度差在(0.9,1.0]之间的分布

Table 4 Distribution of the maximum confidence difference between (0.9, 1.0] for different known and new classes combinations in mixed dataset

已知类 *	新类	已知类 CfD_{max} 分布占比/%	新类 CfD_{max} 分布占比/%
2、14、16	5、6、7、12、13	90.1	7.65
6、8、14、16	3、7、10、11、13	88.6	11.50
3、8、9、16	4、7、10、11、12	78.1	15.20
4、5、7、14、16	3、6、10、11、12	90.5	27.10
2、4、7、12、14	6、9、11、13、15	81.8	31.60
3、9、11、12、15	2、4、6、10、14	80.8	49.20
6、7、10、13、16	2、4、9、12、14	93.6	13.00
1、5、6、7、8、16	3、9、10、11、13	91.7	35.60
1、2、9、11、13、15	3、5、10、14、16	84.1	23.30
5、7、10、11、15、16	1、2、8、9、12	91.5	28.00
1、4、5、7、11、13	2、8、9、10、16	86.6	25.70
2、3、7、8、9、16	4、6、12、13、14	71.4	10.60
1、3、4、5、6、13、15	7、8、9、12、14	74.9	10.50
1、3、4、6、8、13、15	5、7、9、10、16	72.9	10.20
2、3、5、6、8、11、15	1、4、12、14、16	70.0	26.60
2、6、7、8、10、12、14、15	4、9、13、16	78.6	8.32
1、3、5、6、7、9、11、14	2、4、8、12	85.7	25.90
2、3、4、5、6、10、13、15	7、8、9、11	70.4	1.16
2、4、6、7、8、10、11、13、15	9、14、16	73.6	7.23
1、4、6、8、9、10、11、15、16	5、7、12、13	80.4	16.90

* 注:各标签与数据类别的对应关系:1-BitTorrent、2-Email、3-Facebook_audio、4-Facebook_chat、5-Ftp、6-Hangouts_audio、7-Skype_audio、8-Skype_file、9-Skype_video、10-Youtube、11-douyu_480p、12-huya_480p、13-tencent_480p、14-tencent_720p、15-youku_720p、16-douyu_1080p

2.6 未标记样本的筛选

分类器 H_2 选择最大置信度差阈值 $\beta = 0.9$ 对已知类和新类进行分类,若样本的最大置信度差大于阈值,被分类为已知类;若样本的最大置信度差小于阈值,被分类为新类。但新类的最大置信度差也会有部分大于 0.9,对于这部分数据,H_2 会将其误分为已知类。

二分类器 H_1 由已知类和未标记数据中大于最大置信度差阈值的数据进行训练,可以解决 H_2 的误分问题。实际网络中,可以较容易地收集大量未标记流数据[18]。本文尝试从未标记数据中选择负类样本,用以训练 H_1,其目的是为了获得非已知类且大于 CfD_{max} 阈值的负类样本,筛选过程如算法 1 所示。

算法 1 未标记流数据筛选算法

输入 已知类 $z(z_1, z_2, \cdots, z_k)$ 未标记流数据 U 最大置信度差阈值 β

输出 非已知类且最大置信度差 大于 β 的未标记数据子集 M

1.For $i = \{1, 2, \cdots, k\}$ do:

2. 用已知类 z_i 进行训练,得到 OC - SVM$_i$

3.For $i = \{1, 2, \cdots, k\}$ do:

4. 用 OC - SVM$_i$ 对 U 进行分类,得到 U_{i1} 和 U_{i-1}

5. $M1 = U_{11} \cup U_{21} \cup U_{31} \cdots \cup U_{k1}$

6.For m in $M1$:

7. If m only in $U_{11} \parallel U_{21} \parallel U_{31} \cdots \parallel U_{k1}$:

8. $m \in M2$

9. Else:

10. $m \in M3$

11. $M4 = U_{1-1} \cap U_{2-1} \cap U_{3-1} \ldots \cap U_{k-1}$

12. $M5 = M3 \cup M4$

13.RF 输出 $M5$ 中每个样本 m 的置信度集合 $m_t = \{t_1, t_2, \cdots, t_k\}$

14.计算 m_t 中置信度 max 和置信度 min 的差值:$\beta_m = t_{max} - t_{min}$

15.For m in M5:

16. If $\beta_m > \beta$ then:

17. $m \in M$

18. Else:

19. 丢弃 m

注:U_{i1}、U_{i-1}($i = 1, 2, \cdots, k$),分别是OC -SVM$_i$

分类得出的正类和负类。

2.7　基于最大置信度差的已知类和新类划分

分类器 H_2 由已知类数据进行训练，不仅可以通过最大置信度差阈值进行新类检测，还可以通过最大置信度对已知类进行细分类，算法 2 描述了已知类和新类级联分类过程。

算法 2　已知类和新类级联阈值分类算法

输入　混合流数据 X　最大置信度差阈值 β

输出　已知类标签 $z(z_1,z_2,\cdots,z_k)$ 新类 y

1. 对流数据 X 进行特征提取和特征选择，得到输入样本 x

2. H_1 对 x 进行二分类，得到 y_1、x_o，其中 $y_1 \in y$

3. H_2 输出 x_o 中每个样本 l 的置信度集合 $l_t = \{t_1、t_2、\cdots、t_k\}$

4. 计算 l_t 中置信度 max 和置信度 min 的差值：$\beta_l = t_{\max} - t_{\min}$

5. For l in x_o：

6. If $\beta_l > \beta$ then：

7. $l \in z$

8. 找出 l 的置信度集合 l_t 中最大置信度的下标索引 j

9. if $j=1$：$l \in z_1$；if $j=2$：$l \in z_2$；$\cdots$ if $j=k$：$l \in z_k$

10. Else：

11. $l \in y_2$

12. $y = y_1 \cup y_2$

3　实验结果与分析

3.1　评价指标

算法评估包括对分类性能和时间性能的评估。采用 4 种评价指标评估分类性能，分别是开集总体准确率（NA）[19]、查准率（P）、查全率（R）和 $F1$ 测度（$F1$）[20]。NA 是 AKS 和 AUS 关于 λ 的加权平均，这里 AKS 是已知类的分类准确率，AUS 是新类的分类准确率，λ 是样本中已知类所占比例，$0 < \lambda < 1$；TU、FU 分别表示识别正确和错误的新类样本数；P 是预测为正例结果中正确的比例；R 是所有正例样本中被找出的比例；$F1$ 是 P 和 R 的调和平均，具体计算见式（1）～式（6）：

$$NA = \lambda AKS + (1-\lambda)AUS \tag{1}$$

$$AKS = \frac{\sum_{i=1}^{k}(TP_i + TN_i)}{\sum_{i=1}^{k}(TP_i + TN_i + FP_i + FN_i)} \tag{2}$$

$$AUS = \frac{TU}{TU + FU} \tag{3}$$

$$P = \frac{TP}{TP + FP} \tag{4}$$

$$R = \frac{TP}{TP + FN} \tag{5}$$

$$F1 = \frac{2 * P * R}{P + R} \tag{6}$$

其中，对已知类 i 来说，TP_i、TN_i、FP_i、FN_i 分别表示分类正确的正样本数和负样本数、分类错误的正样本数和负样本数；TU、FU 分别表示识别正确和错误的新类样本数；λ 是样本中已知类所占比例，$0 < \lambda < 1$。

3.2　不同阈值对比

本文选定最大置信度差阈值为 0.9，从分类效果进行验证和比较。从混合数据集中随机抽取了一组已知类和新类组合，其中包括 6 个已知类和 6 个新类。不同数据组合和阈值的分类效果见表 5。

表 5　混合数据集上不同阈值分类效果

Table 5　Classification results of different thresholds on mixed dataset

β	已知类			新类			NA
	$F1$	P	R	$F1$	P	R	
0.6	0.822 2	0.774 1	0.876 7	0.686 8	0.828 2	0.586 7	0.770 9
0.7	0.872 2	0.844 6	**0.901 7**	0.822 2	0.886 3	0.766 7	0.869 6
0.8	0.898 4	0.901 9	0.895 0	0.881 0	0.892 3	0.870 0	0.925 2
0.9	**0.931 3**	**0.978 6**	0.888 3	**0.936 2**	**0.897 6**	**0.978 3**	**0.983 6**

由表 5 可知，随着阈值的提高，已知类和新类的 $F1$ 以及 NA 均不断提高。当阈值为 0.9 时，分类效果最佳。对于已知类，当阈值提高时，查准率提升；对于新类，当阈值提高时，查全率提升。

3.3　不同方法对比

将本文方法与 ASG-SVM[6] 在混合数据集上进行分类性能对比实验，结果见表 6。

此外，还将本文方法与 ASG-SVM 进行时间性

能对比,包括训练时间和分类时间的对比,结果见表 7。与 ASG-SVM 相比,本文方法的已知类 $F1$ 提高约 13%,新类 $F1$ 提高约 3%,NA 提高约 5%。本文方法的训练时间和分类时间也表现更佳。

表 6 混合数据集上不同方法分类性能对比

Table 6 Comparison of classification performances of different methods on mixed dataset

方法	已知类			新类			NA
	$F1$	P	R	$F1$	P	R	
本文方法	**0.931 3**	**0.978 6**	**0.888 3**	**0.936 2**	0.897 6	**0.978 3**	**0.983 6**
ASG-SVM	0.796 4	0.792 9	0.800 0	0.910 2	**0.917 1**	0.903 3	0.934 4

表 7 混合数据集上不同方法时间性能对比

Table 7 Comparison of time performances of different methods on mixed dataset ms/样本

方法	训练时间	分类时间
本文方法	**0.076 7**	**0.042 4**
ASG-SVM	31.170 0	1.841 1

4 结束语

为了更好地对未知网络流量进行检测,本文通过对已知类和新类的置信度分布进行分析,提出一种基于置信度信息与级联结构的新类检测方法,并设计了从未标记数据筛选伪负样本的算法。首先,通过二分类器检测高于阈值的新类样本;其次,使用最大置信度差对剩下数据中低于阈值的新类和已知类样本进行区分,并利用最大置信度对已知类进行细分类。在由 2 个真实网络数据集组成的混合数据集上进行了实验验证,总体准确率大于 90%。与现有代表性方法相比,本文方法的已知类和新类 $F1$ 及总体准确率均有显著提升,且训练和分类时间明显减少。下一步工作是探索模型的快速更新,以应对动态出现不同新类实例的检测场景。

参考文献

[1] 汤萍萍. QoS 感知的网络流分类和聚集方法研究[D]. 南京:南京邮电大学,2021.

[2] MU Xin, TING Kaiming, ZHOU Zhihua. Classification under streaming emerging new classes: A solution using completely-random trees[J]. IEEE Transactions on Knowledge and Data Engineering, 2017, 29(8): 1605-1618.

[3] LEE K T, CHIOU C Y, HUANGChunrong. One-class novelty detection via sparse representation with contrastive deep features[C]//International Computer Symposium (ICS). Piscataway: IEEE, 2020: 61-66.

[4] ZAIDI S, LEE C G. One-class classification based bug triage system to assign a newly added developer[C]//International Conference on Information Networking (ICOIN). Piscataway: IEEE, 2021: 738-741.

[5] 刘欢,郑庆华,罗敏楠,等. 基于跨域对抗学习的零样本分类[J]. 计算机研究与发展,2019,56(12):2521-2535.

[6] YANG Yu, QU Weiyang, LI Nan, et al. Open category classification by adversarial sample generation[C]//International Joint Conference on Artificial Intelligence(IJCAI). San Francisco: Morgan Kaufmann, 2017: 3357-3363.

[7] GENG Chuanxing, HUANG Shengjun, CHEN Songcan. Recent advances in open set recognition: A survey[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021, 43(10): 3614-3631.

[8] SCHEIRER W J, ANDERSON D RR, SAPKOTA A, et al. Toward open set recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2013, 35(7): 1757-1772.

[9] SCHEIRER W J, JAIN L P, BOULT T E. Probability models for open set recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2014, 36(11): 2317-2324.

[10] SCHERREIK M D, RIGLING B D. Open set recognition for automatic target classification with rejection[J]. IEEE Transactions on Aerospace and Electronic Systems, 2016, 52(2): 632-642.

[11] NEAL L, OLSON M, FERN X, et al. Open set learning with counterfactualimages[C]//Proceedings of the 15th European Conference on Computer Vision. Berlin: Springer Verlag, 2018: 620-635.

[12] YANG Yang, HOU Chunping, LANG Yue, et al. Open-set human activity recognition based on micro-Doppler signatures[J]. Pattern Recognition, 2019,85: 60-69.

[13] VARETO R, SILVA S, COSTA F, et al. Towards open-set face recognition using hashing functions[C]//IEEE International Joint Conference on Biometrics. Lyon, France :IEEE, 2018: 634-641.

[14] NERIA M A C, JUNIOR P R M, ROCHA A, et al. Data-fusion techniques for open-set recognition problems[J]. IEEE Access, 2018(6):21242-21265.

[15] DONG Hanze, FU Yanwei, SIGAL L, et al. Learning to separate domains in generalized zero-shot and open set learning: A probabilistic perspective[J]. arXiv preprint arXiv: 1810.07368, 2018.

[16] SABER A, FERGANI B, ABBAS M. Encrypted traffic classification: Combining over-and under-sampling through a PCA-SVM[C]//Proceedings of the 3rd International Conference on Pattern Analysis and Intelligent Systems (PAIS). Tebessa, Algeria:IEEE, 2018: 1-5.

[17] 项阳,董育宁,魏昕. 一种基于机器学习的网络流早期分类方法[J]. 南京邮电大学学报(自然科学版),2022,42(4):96-104.

[18] 郭翔宇. 利用未标记数据的机器学习方法研究[D]. 南京:南京大学,2017.

[19] 高菲,杨柳,李晖. 开放集识别研究综述[J]. 南京大学学报(自然科学版),2022,58(1):115-134.

[20] 刘会霞,董育宁,邱晓晖. 基于相关性特征选择和深度学习的网络流分类[J]. 南京邮电大学学报(自然科学版),2022,42(4):75-84.