吕超, 董育宁, 邱晓晖. 一种基于 ELM 算法的在线学习模型[J]. 智能计算机与应用,2024,14(6):110-118. DOI:10.20169/j.issn.2095-2163.240615

一种基于 ELM 算法的在线学习模型

吕 超, 董育宁, 邱晓晖
(南京邮电大学 通信与信息工程学院, 南京 210003)

摘 要: 网络应用程序的多样化对网络流量分类提出了新的挑战。如何在变化的环境中准确地识别已知类和新类流量,然后实现模型在线更新,最后将新类纳入已知类范畴成为了研究的要点。针对这一问题,本文提出了一种基于极限学习机(Extreme Learning Machine, ELM)的在线学习模型,使用基于 ELM 算法的距离度量选择辅助训练样本,根据距离度量阈值进行新类检测,采用串联识别新类的二分类器的方式包含新的流量类别,当串联的分类器数量达到设定值时重新训练模型。在真实网络流数据集上的测试结果显示,本文方法已知类 *F*1 和开集总体准确率 *NA* 均能达到 0.9 以上。与代表性文献方法相比,在分类性能和时间性能方面均有更好的表现。

关键词: 极限学习机; 开集流识别; 新类检测; 辅助训练; 在线学习
中图分类号: TP391 **文献标志码**: A **文章编号**: 2095-2163(2024)06-0110-09

An online study model based on ELM algorithm

LÜ Chao, DONG Yuning, QIU Xiaohui
(School of Communications and Information Engineering, Nanjing University of Posts and Telecommunications, Nanjing 210003, China)

Abstract: The diversity of network applications poses new challenges to network traffic classification. How to accurately identify the known class and the new class traffic in the changing environment, then realize the online model update, finally include the new class into the known class category has become the key point of research. To solve this problem, an online learning model based on Extreme Learning Machine (ELM) algorithm is proposed in this paper. The distance measurement based on ELM algorithm is used to select auxiliary training samples, and the new class detection is carried out according to the distance measurement threshold. The binary classifier of the new class is identified in series to include the new traffic class, and the model is retrained when the number of series classifiers reaches the set value. The test results on real network flow data sets show that the known *F*1 scores and open set overall accuracy *NA* of the proposed method can reach above 0.9. Compared with representative literature method, it has better performance in the classification and time consumption.

Key words: Extreme Learning Machine; open-set flow recognition; new class detection; auxiliary training; online study

0 引 言

近些年,互联网的迅速发展使得数据量不断增长,为网络流分类(Network Traffic Classification, NTC)带来了新的困难和挑战[1-2],尤其是网络应用程序层出不穷,需要检测的流量类别越来越多,对于在网络流分类过程中包含未知类的流识别问题,统称为开集流识别(Open Set Flow Recognition, OSFR)[3]问题。

OSFR 和其它的开放世界识别(Open Set Recognition, OSR)[4]问题一样,希望模型能够自动完成 3 个过程:首先是新类检测问题,即如何将已知类和新类样本加以有效区分,然后对识别出的新类进行标记并将其加入到已知类中,最后是增量学习过程,根据新的已知类集合更新模型,将新类纳入已知类范畴中。

针对 OSFR 中的 3 个过程,本文提出了一种基于极限学习机(Extreme Learning Machine, ELM)[5]

作者简介: 吕 超(1998-),男,硕士研究生,主要研究方向:网络流量分类; 邱晓晖(1968-),女,博士,教授,主要研究方向:智能信号处理,图像处理,模式识别。
通讯作者: 董育宁(1955-),男,博士,教授,主要研究方向:无线通信网络,图像处理。Email:19900011@njupt.edu.cn

收稿日期: 2023-04-12

的网络流量在线学习模型(Online Study Model for Network Traffic Classification based on ELM Algorithm,OSNTC-ELM)。ELM 结构与单隐层前馈神经网络相同,训练时采用随机的输入层权值和偏差,而不是传统神经网络中的基于梯度的算法,具有逼近能力强[6]、学习速度快[7]、容易实现和最小的人工干预等特点[8],不仅能够满足网络流量在线分类的要求,还可以基于距离度量阈值进行新类检测。在单新类出现的假设下设计出一种模型更新方法,本文的主要贡献如下:

(1)提出了一种基于 ELM 的距离度量进行新类检测,具备一定的新类检测效果。

(2)提出了一种基于 ELM 距离度量的辅助训练样本的选择方法,可以有效提升新类检测能力。

(3)设计了一种模型更新的方法,采用串联新的二分类器方式,有效降低了模型更新时间,避免每次添加新的已知类时重新训练模型,只有在满足一定条件的时候才会重新训练模型。

(4)在 2 个真实数据集上验证了方法的有效性。相比于已有方法,在分类性能和时间性能上具有一定优势。

论文的其余部分安排如下。第 1 节介绍了相关的 OSFR 方法;第 2 节详细叙述本文方法;第 3 节给出本文方法与文献方法比较结果;第 4 节是总结。

1　相关工作

近年来,OSFR 受到越来越多的关注,也涌现出了很多的解决方法。一类样本分类问题[9-11]是专门针对测试集中出现新类而提出的方法,基本思想是基于某些度量拒绝不属于已知类的样本,和异常检测方法类似,但是却不能对已知类进行分类。一类样本分类问题为 OSFR 开拓了新思路,基于这一思想,主流的方法分成了两大类,分别是:基于机器学习(Machine Learning, ML)[12-13]和基于深度学习(Deep Learning,DL)[14-15]的方法。

基于 ML 的方法主要是基于支持向量机(Support Vector Machine,SVM)[16]的方法, Scheirer 等学者[17]提出了一种“1-vs-set”机制,通过在核空间构建和 SVM 超平面平行的另一超平面,将已知类限制在 2 个超平面之间,想法相似的还有 Cevikalp[18]提出使用最佳拟合平面算法来使每个超平面接近某一类样本。Cevikalp 等学者[19]提出一组准线性多面体二次曲线判别算子,这类线性算子可以通过非对称分类器为某类正样本生成更加紧凑、约束良好的决策边界。这些方法在新类检测方面都有较好的表现,但是已知类性能不佳。

其它还有基于 HDP(Hierarchical Dirichlet Process)[20-21]的 CD-OSR(Collective Decision-based OSR)模型,不使用阈值,依靠自动聚类为新类留出空间,新类样本会自动聚成一类。此方法性能较为均衡,但是会消耗更多的时间用于训练,不满足在线流分类的要求。

基于 DL 的方法主要是基于生成对抗网络(Generative Adversarial Networks, GANs)[22-24]的方法, Neal 等学者[25]借助 GANs 对训练集样本进行扩充,生成的样本很接近已知类,但不属于任何已知类。Yang 等学者[26]将其与 SVM 结合起来,提出了对抗样本生成(Adversarial Sample Generation, ASG)策略,先使用 GANS 为已知类[10-11]生成对应的负类样本,然后为每个已知类训练了一个 SVM 二分类器,测试时,如果一个样本被所有分类器判定为负类,那么判定其为新类样本。实验结果表明此方法在识别新类方面效果显著。

2　本文方法

在闭集流分类问题中,ELM 分类器在输出的多个预测结果中,选择其中最接近 1(原型样本的输出结果为 1)的结果对应的类别作为预测标签。在 OSFR 中,这种方法会将新类判定为某一个已知类,不过,已知类样本和新类样本的预测结果有明显差异。为了方便描述,定义变量 $d_{\min}$, 表示预测结果与原型样本之间的距离,可由式(1)来确定:

$$d_{\min} = \min | res - 1 | \tag{1}$$

其中, res 是分类器的预测结果。那么闭集问题中,预测标签就是 $d_{\min}$ 对应的类别。

研究发现已知类的 $d_{\min}$ 分布往往比较集中,且值较小,而新类的 $d_{\min}$ 分布比较分散,且值都比较大,两者分布具有明显区别。基于这一发现,寻找一个合适的阈值 β 来进行新类检测是一个可行的方法。不过,部分新类样本的 $d_{\min}$ 也会小于 β, 为了解决这一问题,提出了基于 ELM 距离度量的选择辅助训练样本的方法,基本思想是选择和已知类相似、但不属于已知类的样本作为辅助训练样本,具体方法是在无标签数据集中筛选 $\beta < d_{\min} < \beta + \varepsilon$ 的样本作为辅助训练样本。

图 1 给出了本文方法的基本框图。经过数据采集和特征提取两个阶段后,将特征输入二分类器 H_0, H_0 使用已知类和辅助训练样本训练,已知类为

正样本,辅助训练样本为负样本。H_0 判决规则如下:判定为正类(P)且$d_{\min} \leqslant \beta$的样本为正样本,其余为负样本。正样本进入多分类器 H_E 进行细分类,负样本视作新类样本进入新类处理模块。

图 2 给出了新类处理模块的基本框图。首先收集足够的新类样本,然后在无标签数据集中筛选辅助训练样本,最后新类样本作为正样本,辅助训练样本作为负样本,训练识别该新类的二分类器 H_{ni},H_{ni} 判决规则和 H_0 相同。

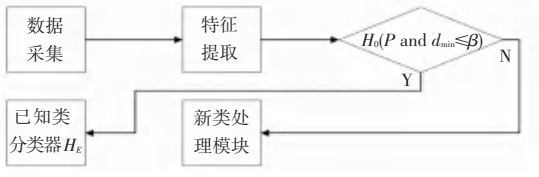

图 1 基于 ELM 算法的在线学习方法

Fig. 1 Online learning method based on ELM algorithm

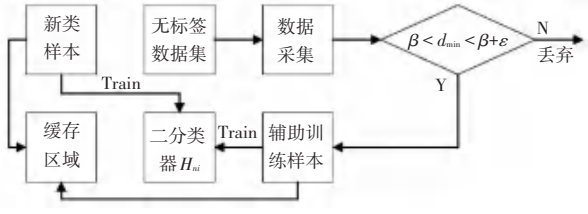

图 2 新类处理模块流程

Fig. 2 Basic structure of the new class samples processing module

在线学习是一个动态的过程,本文的更新思路是将新训练的 H_{ni} 依次串联在 H_0 与新类处理模块之间,当初始已知类数目 m 与缓存中类别数 n 相等时,进行整体更新,为了更好地说明本文的更新方法,图 3 中展示了 $m = 3$ 时的更新过程。

初始时刻 $T0$,初始已知类数目 $m = 3$,缓存中类别数目 $n = 0$,可识别类别数目 $k = 3$,测试集样本中有 3 个已知类和 1 个新类,测试样本经过 H_0 判别后,正样本进入 H_{E3}、判别后输出预测标签,负样本进入新类处理模块、训练出二分类器 H_{n1},将其串联在 H_0 和新类处理模块之间,完成第一次更新。

$T1$ 时刻,$m = 3$,$n = 1$,$k = 4$,测试集样本中有 4 个已知类和 1 个新类,先经过 H_0 判别,正样本进入 H_{E3}、判别后输出预测标签,负样本进入 H_{n1} 判别,正样本获得预测标签,负样本进入新类处理模块,训练出二分类器 H_{n2},串联在 H_{n1} 后,完成第二次更新。

$T2$ 时刻,$m = 3$,$n = 2$,$k = 5$,具体流程和 $T0$、$T1$ 时刻类似。

$T3$ 时刻,$m = 3$,$n = 3$,$k = 6$,此时 $m = n$,满足整体更新的要求,使用 3 个初始已知类和缓存区域的样本,训练 H_0^* 和 H_{E6} 替换 H_0 和 H_{E3},进一步清空 H_0^* 和新类处理模块间的所有二分类器,完成整体更新,实现了增量学习。

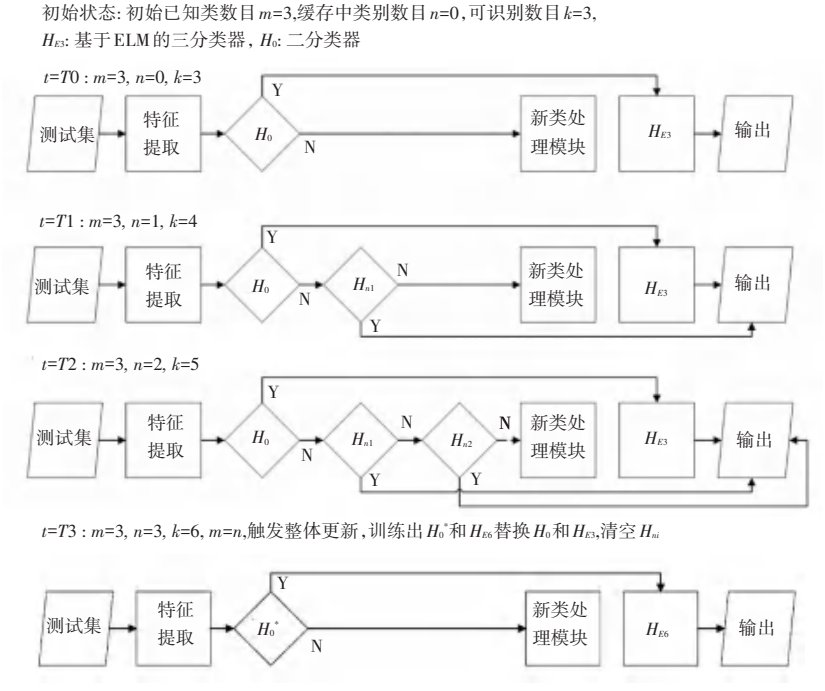

图 3 $m = 3$ 的更新流程

Fig. 3 Update process when $m = 3$

2.1　数据集

实验使用 2 个真实网络流数据集：南邮数据集（NY）和 ISCX non－VPN（ISCX）[27]；其中，NYD 是 2020 年在南京邮电大学校园网使用 Wireshark[28] 软件采集得到，ISCX 是公开数据集，表 1 和表 2 给出数据集的具体信息。

表 1　NY 数据集

Table 1　NY data set

流类型	应用	样本数量
视频点播	Tencent、Bilibili、Youku	2 000×3
网络音乐	Cloud_music、Kugou、QQ_music	2 000×3
视频直播	Douyu、Huya	2 000×2
视频通话	Tencent_meeting	2 000
网络聊天	Wechat	2 000

表 2　ISCX 数据集

Table 2　ISCX data set

流类型	应用	样本数量
语音通话	Facebook、Hangsout、Skype	2 000×3
视频	BitTorrent、Skype、YouTube	2 000×2
下载	Ftp、skype_file	2 000×2
IP 语音	VoipBuster	2 000
文字聊天	Facebook	2 000

2.2　d_{min} 分布与阈值 β

阈值 β 是新类检测的关键，而 β 依靠样本的 d_{min} 分布确定。为了确定最佳的阈值 β，在 ISCX 数据集中随机选择 5 个类作为已知类，剩余 5 个类作为新类，用已知类训练多分类器。测试时，每个类别使用 500 个样本，按照已知类和新类统计对应 d_{min} 分布，结果如图 4 所示。

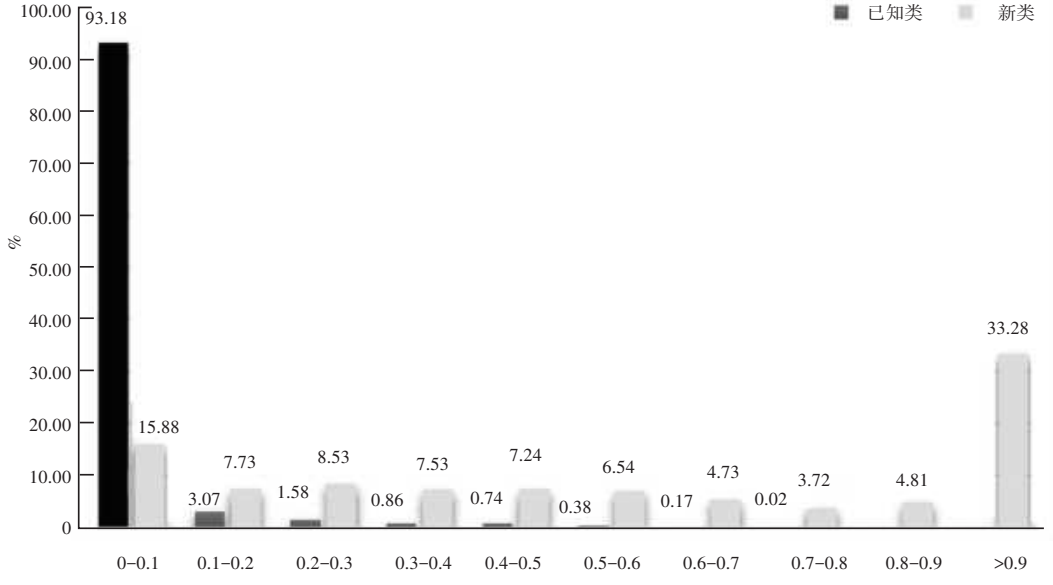

图 4　ISCX 数据集上已知类和新类的 d_{min} 分布

Fig. 4　d_{min} distributions of known and new classes on the ISCX dataset

由图 4 可知，已知类 d_{min} 集中分布在[0,0.1]的区间内，新类样本分布则相对分散，约有 16% 样本分布在[0,0.1]内，[0.1,0.9]的区间内分布比较均衡，但比例都不高，d_{min} 大于 0.9 的样本占最大比例。图 4 的结果说明了使用阈值进行新类检测是可行的，阈值 β 设为 0.1 就可以进行新类检测，可以识别出 93% 左右的已知类和 84% 左右的新类样本。不过仍然有约 16% 的新类样本会被判定为已知类，对于已知类分类性能有较大的影响。

2.3　辅助训练样本的选择

为了解决 2.2 节中部分新类样本会被判定为已知类的问题，本文提出了使用辅助训练样本的思路，选出和已知类相似但不属于已知类的样本作为辅助训练样本，具体方法是在无标签数据集中收集 $d_{min} \in (\beta, \beta+\varepsilon]$ 的样本。为了满足和已知类相似的标准，ε 应该设置得比较小。为了验证该思路的有效性，使用和 2.2 节中相同的已知类和新类，β 设为 0.1，ε 设为 0.01，在无标签数据集中筛选对应辅助训练样本，使用已知类和辅助训练样本训练一个二分类器。分别使用已知类和新类样本进行测试，统计对应 d_{min}，为了方便统计，如果一个样本判定为正类，d_{min} 按照区间进行统计，判定为负类，令其

$d_{\min}$ =1, 然后进行统计。结果如图 5 所示。

和图 4 的结果相比,已知类 $d_{\min}$ 仍然集中分布在[0,0.1]的区间内,不过比例下降了 3%左右,但是分布在[0,0.1]的区间内的新类样本的比例有了明显的下降,下降了 10%左右,说明了辅助训练样本的有效性。

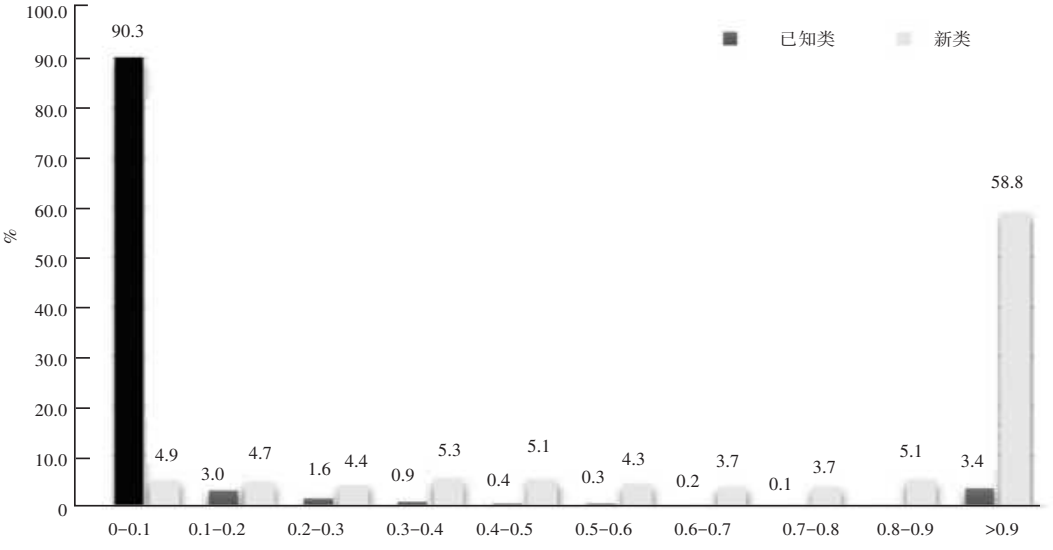

图 5 使用辅助训练样本后的$d_{\min}$分布

Fig. 5 $d_{\min}$ distribution after using auxiliary training samples

2.4 基于 ELM 的在线学习算法

本文方法的算法分成新类检测兼已知类识别和模型更新这 2 个过程。算法 1 给出了新类检测兼已知类识别的算法描述。算法 2 描述模型更新的整体流程。

算法 1 新类检测及已知类识别

输入 $ELM-set(H_0、H_E)$; β -分类阈值; D -测试集; D_{new} - 新类样本集合

输出 y - 预测标签($k_1、k_2、k_m、n$)

for each x in D do:
  $d_{\min} = H_0(x)$
  if pre = postive and $d_{\min} \leqslant \beta$:
    $y \leftarrow H_E(x)$
  else:
    $y \leftarrow n$
    $D_{\text{new}} \leftarrow D_{\text{new}} \cup \{x\}$
end for

算法 2 模型更新

输入 m -已知类数量; $size$ -新类训练所需数量; β -分类阈值; ε -辅助选择阈值; $ELM\_new-set$; $ELM(H_d)$; D_{help} - 辅助训练样本集合; D_{new} - 新类样本集合; $D_{\text{no\_label}}$ - 无标签数据集; $D_{\text{k\_train}}$ - 已知类训练集

输出 $ELM-H_0^*$; $ELM-H_E^*$

while $sizeof(D_{\text{help}}) = size$:
  for each x in $D_{\text{no\_label}}$:
    $d_{\min} = H_d(x)$
    if $d_{\min} \in (\beta, \beta+\varepsilon]$
      $D_{\text{help}} \leftarrow D_{\text{help}} \cup \{x\}$
  end for
end while
use D_{new} and D_{help} train a new ELM- H_{ni}
$ELM\_new-set \leftarrow ELM\_new-set \cup H_{ni}$
when $sizeof(ELM\_new-set) == m$:
  use $D_{\text{k\_train}}、D_{\text{new}}$ and D_{help} train ELM H_0^* and H_E^*
  replace H_0 and H_E with H_0^* and H_E^*
  clear H_{ni}

3 实验结果与分析

3.1 评估指标

实验评估指标包括分类准确性和时间性能评估。分类准确性包含查准率(P)、查全率(R)、$F1$ 测度($F1\_score$) 以及开集总体准确率(NA)。分别可由式(2) ~ 式(7) 计算求得:

$$P = \frac{TP}{TP + FP} \tag{2}$$

$$R = \frac{TP}{TP + FN} \tag{3}$$

$$F1 = \frac{2 \times P \times R}{P + R} \tag{4}$$

$$NA = \lambda AKS + (1 - \lambda) AUS \tag{5}$$

$$AKS = \frac{\sum_{i=1}^{m} (TP_i + TN_i)}{\sum_{i=1}^{m} (TP_i + TN_i + FP_i + FN_i)} \tag{6}$$

$$AUS = \frac{TU}{TU + FU} \tag{7}$$

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 表示测试样本中已知类占比。时间性能包括训练时间、在线分类时间和模型更新时间三个方面。

3.2　实验场景和参数设置

实验环境为 Lenovo Legion Y70002021 笔记本，操作系统为 Window 11，CPU 为 11th Gen intel Core i5-11400H @ 2.70 GHz，运行内存 16 GB。使用 Python 实现 ELM 算法，ELM 神经元个数选择 120，隐层神经元输入权重为 (-1,1) 内随机数，隐层神经元偏置为(-0.4,0.4)内随机数，新类检测阈值 β 设为 0.1，辅助训练样本筛选阈值 ε 设为 0.01，新类样本数量达到 500 时训练二分类器。实验采用五折交叉验证。

3.3　实验结果分析

3.3.1　OSNTC-ELM 分类性能

为了验证本文方法的分类性能，按照图 3 的流程进行实验，测试分成 4 个阶段，分别在 2 个数据集上各选择 7 个类别进行实验，其中 3 个作为初始已知类，4 个作为新类，依次增加，测试过程中各阶段信息见表 3。统计所有类别的具体性能，结果见表 4 和表 5。

表 3　测试阶段类别信息

Table 3　Class information of the test phase

Period	known class	new class
*T*0	*c*1,*c*2,*c*3	*c*4
*T*1	*c*1,*c*2,*c*3,*c*4,	*c*5
*T*2	*c*1,*c*2,*c*3,*c*4,*c*5,	*c*6
*T*3	*c*1,*c*2,*c*3,*c*4,*c*5,*c*6	*c*7

由表 4、表 5 可知，本文的方法在 2 个数据集上都取得不错的分类性能，不过随着更新次数的增加，分类的性能会有所下降，在增加二分类器更新中，下降得并不明显，在进行整体更新后，*NA* 会有 2% ~ 4% 的性能下降。

表 4　ISCX 数据集上性能表现

Table 4　Performance on ISCX data set

ISCX	Known/New	*P*	*R*/*AUS*	*F*1	*NA*
*T*0	Ftp_K	0.978	0.988	0.983	0.978
	Hangsout_K	0.944	0.974	0.959	0.978
	SkypeAudio_K	1	0.956	0.978	0.978
	SkyprVideo_N	-	0.963	-	0.978
*T*1	Ftp_K	0.988	0.998	0.993	0.974
	Hangsout_K	0.970	0.968	0.969	0.974
	SkypeAudio_K	0.984	0.954	0.969	0.974
	SkyprVideo_K	0.956	0.920	0.938	0.974
	VoipBuster_N	-	0.946	-	0.974
*T*2	Ftp_K	0.942	0.980	0.961	0.969
	Hangsout_K	0.962	0.972	0.967	0.969
	SkypeAudio_K	0.958	0.956	0.957	0.969
	SkyprVideo_K	0.915	0.878	0.898	0.969
	VoipBuster_K	0.975	0.874	0.922	0.969
	Youtobe_N	-	0.926	-	0.969
*T*3	Ftp_K	0.998	0.918	0.956	0.956
	Hangsout_K	0.993	0.904	0.947	0.956
	SkypeAudio_K	0.996	0.982	0.989	0.956
	SkyprVideo_K	0.991	0.872	0.928	0.956
	VoipBuster_K	0.925	0.964	0.965	0.956
	Youtobe_K	0.998	0.970	0.984	0.956
	Facebook_N	-	0.903	-	0.956

表 5　NY 数据集上性能表现

Table 5　Performance on NY data set

NYD	Known/New	*P*	*R*/*AUS*	*F*1	*NA*
*T*0	Game_K	0.963	0.986	0.974	0.963
	Huya_K	0.992	0.960	0.976	0.963
	KugouMusic_K	0.994	0.956	0.975	0.963
	QQMusic_N	-	0.951	-	0.963
*T*1	Game_K	0.938	0.998	0.967	0.957
	Huya_K	0.988	0.972	0.980	0.957
	KugouMusic_K	0.946	0.946	0.946	0.957
	QQMusic_K	0.897	0.878	0.887	0.957
	TxMeeting _N	-	0.919	-	0.957
*T*2	Game_K	0.958	0.992	0.974	0.955
	Huya_K	0.996	0.976	0.986	0.955
	KugouMusic_K	0.932	0.956	0.944	0.955
	QQMusic_K	0.894	0.846	0.869	0.955
	TxMeeting _N	0.949	0.888	0.917	0.955
	Wechat_N	-	0.956	-	0.955
*T*3	Game_K	1	0.976	0.984	0.921
	Huya_K	1	0.958	0.976	0.921
	KugouMusic_K	0.998	0.876	0.933	0.921
	QQMusic_K	1	0.870	0.930	0.921
	TxMeeting _N	0.996	0.904	0.949	0.921
	Wechat_K	1	0.974	0.987	0.921
	Douyu_N	-	0.931	-	0.921

整体更新前性能下降不明显的原因是后训练出来的 H_{ni} 中包含部分的初始已知类样本特征信息，而大部分初始已知类样本会被隔离在 H_0 处，少量进

入 H_{ni} 的样本往往也会因为 $d_{\min} > \beta$ 而被判定为新类。但是整体更新时,会使用缓存中不纯的样本重新训练分类器,所以会出现性能下降的现象。

3.3.2 与代表方法的比较

将本文方法与 CD-OSR 方法[20] 和 ASG 方法[26] 进行比较,分别在 2 个数据集上按照图 3 流程进行测试,统计 *T*0 和 *T*3 阶段分类性能。在 ASG 方法中,每次更新并联新训练的 SVM 二分类器。在 CD-ORS 方法中,每次更新使用已知类和新类样本进行重训练,实验结果见表 6、表 7。

表 6 *T*0 阶段 ASG、CD-OSR、OSNTC-ELM 性能

Table 6 Performance of the ASG、CD-OSR and OSNTC-ELM in phase *T*0

Dataset	Method		*P*	*R/AUS*	*F*1	*NA*
NY	ASG	K	0.887	0.846	0.866	0.921
		N	-	0.957	-	0.921
	CD-OSR	K	0.907	0.804	0.852	0.857
		N	-	0.845	-	0.857
	OSNTC-ELM	K	**0.938**	**0.924**	**0.931**	**0.963**
		N	-	**0.976**	**-**	**0.963**
ISCX	ASG	K	0.892	0.904	0.898	0.927
		N	-	0.937	-	0.927
	CD-OSR	K	0.914	0.805	0.856	0.875
		N	-	0.877	-	0.875
	OSNTC-ELM	K	**0.967**	**0.965**	**0.966**	**0.983**
		N	-	**0.980**	**-**	**0.983**

表 7 *T*3 阶段 ASG、CD-OSR、OSNTC-ELM 性能

Table 7 Performance of the ASG、CD-OSR and OSNTC-ELM in phase *T*3

Dataset	Method		*P*	*R/AUS*	*F*1	*NA*
NY	ASG	K	0.754	0.719	0.736	0.775
		N	-	0.813	-	0.775
	CD-OSR	K	0.771	0.683	0.725	0.685
		N	-	0.718	-	0.685
	OSNTC-ELM	K	**0.919**	**0.906**	**0.912**	**0.925**
		N	-	**0.927**	**-**	**0.925**
ISCX	ASG	K	0.758	0.768	0.763	0.741
		N	-	0.796	-	0.741
	CD-OSR	K	0.777	0.684	0.728	0.701
		N	-	0.745	-	0.701
	OSNTC-ELM	K	**0.938**	**0.936**	**0.937**	**0.951**
		N	-	**0.931**	**-**	**0.951**
		N	-	**0.937**	**-**	**0.951**

\

由表 6 可知,在进行更新前,本文方法稍优于对比方法,与 ASG 方法相比,本文方法已知类 *F*1 平均高 6.8%, *NA* 平均高 5.3%。ASG 方法在已知类分类方面表现不佳,容易把已知类样本划分为新类,所以已知类 *Recall* 较低,因为对抗网络生成的负类样本与已知类的边缘样本高度相似,利用这些样本训练的二分类器容易将已知类的边缘样本判定为新类。不过也因具有优秀的新类检测能力,在 2 个数据集上 *AUS* 均能达到 0.9 以上。

与 CD-OSR 方法相比,本文方法已知类 *F*1 平均高 9.2%, *NA* 平均高 10.2%。基于聚类的算法不使用标签信息,采用的是无监督的训练方式,在训练

过程中就可能会将不同类别的样本聚类到同一个类别，所以在已知类和新类两方面都有所欠缺。

由表 7 可知，与整体更新前相比，3 种方法的分类性能都有所下降，本文的方法下降的程度比较低，更新后已知类和新类检测性能下降在 2%～4%，ASG 方法和 CD-OSR 方法则下降明显，2 种方法均有 15%以上的性能下降，因为使用纯度不足的新类样本进行训练带来的错误在更新过程中不断累积，导致更新后性能下降严重。

表 8 是本文方法与对比方法时间性能的对比。时间性能包含 3 个方面：训练时间、在线分类时间和模型更新时间。在训练阶段，ELM 因为单隐层的结构，使用的神经元个数较少，所以训练速度比较快。CD-OSR 方法是基于聚类算法，训练时间较长。ASG 需要为每个类训练对应的二分类器，所以时间消耗最长。

在线测试阶段，因为本文方法采用的是串联结构，所以仅有一小部分样本（新类和已知类异常样本）需要经过所有训练器，减少了时间消耗。而 ASG 采用的是并联的结构，所有的分类器都需要对样本进行计算，并且随着更新分类器数量会增加，分类时间也会增加。CD-OSR 方法的分类速度较快，所以消耗的时间不多。

在模型更新阶段，ELM 需要选择辅助训练样本，所以更新时间比训练时间长。ASG 方法则是和训练时间相差不多，因为只需要训练一个新的二分类器。CD-OSR 方法要进行重训练，为此更新时间和训练时间相比，略有增加。三者相比，本文方法更新时间还是最短的。

表 8　ASG、CD-OSR、OSNTC-ELM 各个阶段时间对比

Table 8　Time comparison of ASG、CD-OSR、and OSNTC-ELM in different phases

ms/样本

Dataset	Method	训练时间	测试时间	模型更新时间
NY	OSNTC-ELM	**0.173**	**0.076**	**0.284**
	ASG	27.334	1.641	27.964
	CD-OSR	6.362	0.391	7.227
ISCX	OSNTC-ELM	**0.161**	**0.062**	**0.269**
	ASG	24.573	1.241	24.359
	CD-OSR	5.384	0.353	6.172

4　结束语

本文提出了一种基于 ELM 距离度量的新类检测方法，并在此基础上设计出一种模型更新的方式。在结合辅助训练样本的情况下，2 个数据集的实验结果表明，更新前后对于新类的检测能力均能达到 0.9 以上，对于已知类而言，$F1$ 也能达到 0.9 以上。虽然更新前后整体性能 NA 有所下降，不过下降的幅度比较小。与对比方法相比，本文方法的已知类 $F1$ 及开集总体准确率 NA 更高，并且训练时间、分类时间以及模型更新时间更短。

不过本文的方法还存在一定的局限性。在 2 个数据集上阈值都是 0.1 时性能最佳，但是不排除最佳阈值会受到不同数据集的影响。在后续的工作中，可以考虑对缓存区中的新类样本做更进一步的提纯，比如设置阈值来选择用于训练的样本，或对更新后的模型参数进行调整，优化整体性能等。

参考文献

[1] MASUD M M, WOOLAM C, GAO Jing, et al. Facing the reality of data stream classification: Coping with scarcity of labeled data [J]. Knowledge and Information Systems, 2012, 33(1): 213-244.

[2] READ J, BIFET A, HOLMES G, et al. Scalable and efficient multi-label classification for evolving data streams [J]. Machine Learning, 2012, 88: 243-272.

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

[4] BENDALE A, BOULT T. Towards open world recognition[C]// Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA: IEEE, 2015: 1893-1902.

[5] YANG Zhe, LONG Jianyu, ZI Yanyang. et al. Incremental novelty identification from initially one-class learning to unknown abnormality classification [J]. IEEE Transactions on Industrial Electronics, 2022, 69(7): 7394-7404.

[6] CHEN Yuanyuan, WANG Zhibin. Cross components calibration transfer of NIR spectroscopy model through PCA and weighted ELM based TrAdaBoost algorithm [J]. Chemometrics & Intelligent Laboratory Systems, 2019, 192(15): 103824-103843.

[7] LIU Tan, XU Tongyu, YU Fenghua. A method combining ELM and PLSR (ELM-P) for estimating chlorophyll content in rice with feature bands extracted by an improved ant colony optimization algorithm [J]. Computers & Electronics in Agriculture, 2021, 186(2): 106-110.

[8] YE Ansheng, ZHOU Xiangbing, MIAO Fang. Innovative hyperspectral image classification approach using optimized CNN and ELM [J]. Electronics-Switz, 2022, 11(5): 775-778.

[9] JAIN L P, SCHEIRER W J, BOULT T E. Multi- class open set recognition using probability of inclusion[C]//Proceedings of the 13th European Conference on Computer Vision. Zurich, Switzerland: Springer, 2014: 393-409.

[10] BODESHEIM P, FRETAG A, RODNER E, et al. Kernel null space methods for novelty detection [C]//Proceedings of 2013 IEEE Conference on Computer Vision and Pattern Recognition. Portland, USA: IEEE, 2013: 3374-3381.

[11] TAX D M J, DUIN R P W. Growing a multi-classclassifier with a

reject option[J]. Pattern Recognition Letters,2008,29(10):1565-1570.

[12]CHENG Guang, WANG Song.Traffic classification based on port connection pattern [C]//International Conference on Computer Science and Service System (CSSS). Nanjing, China: IEEE, 2011:914-917.

[13]KWON J, JUNG D, PARK H. Traffic data classification using machine learning algorithms in SDN networks[C]// International Conference on Information and Communication Technology Convergence (ICTC). Jeju, Republic of Korea: IEEE, 2020: 1031-1033.

[14]YANG Lixuan, FINAMORE A, FENG Jun, et al. Deep learning and zero-day traffic classification: Lessons learned from a commercial-grade dataset[J]. IEEE Transactions on Network and Service Management, 2021, 18(4): 4103-4118.

[15]FINSTE R, BUSCH M, RICHTER C, et al.A survey of payload-based traffic classification approaches[J]. IEEE Communications Surveys & Tutorials,2014,16(2): 1135-1156.

[16]CORTES C, VAPNIK V. Support-vector networks[J].Machine Learning, 1995, 20(3): 273-297.

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

[18]CEVIKALP H. Best fitting hyperplanes for classification[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(6):1076-1088.

[19]CEVIKALP H, TRIGGS B. Polyhedral conic classifiers for visual object detection and classification[C]// Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu,USA:IEEE,2017:4114-4122.

[20]GENG Chuanxing, CHEN Songcan. Collective decision for open set recognition[J]. IEEE Transactions on Knowledge and Data Engineering, 2022, 34(1): 192-204.

[21]GERSHMAN S J, BLEI D M. A Tutorial on Bayesian nonparametric models[J]. Journal of Mathematical Psychology, 2012, 56(1): 1-12.

[22]GOODFELLOW I, POUGET-ABADIE J, MIRZAM, et al. Generative Adversarial Networks[J]. Communications of the ACM, 2020, 63(11): 139-144.

[23]RAMESH A, PAVLOV M, GOH G, et al. Zero-shottextto-image generation [C]//International Conference on Machine Learning. PMLR, 2021: 8821-8831.

[24]BERMANO A H, GAL R, ALALUF Y, et al. State-of-the-art in the architecture, methods and applications of StyleGAN[J]. arXiv preprint arXiv: 2202.14020,2022.

[25]NEAL L, OLSON M, FERN X, et al. Open set learning with counterfactual images [C]//Proceedings of the European Conference on 15th European Conference on Computer Vision. Munich, Germany: Springer, 2018:620-635.

[26]YANG Yu, QU Weiyang, NAN L, et al. Open category classification by adversarial sample generation[C]//International Joint Conference on Artificial Intelligence (IJCAI). Melbourne, Australia: AAAI, 2017: 3357-3363.

[27]LASHKARI A H, GIL G D, MAMUN M, et al. Characterization oftor traffic using time based features [C]//International Conference on Information Systems Security & Privacy (ICISSP). Porto, Portugal: dblp, 2017: 253-262.

[28]DAS R, TUNA G. Packet tracing and analysis of network cameras withWireshark [C]//5th International Symposium on Digital Forensic and Security (ISDFS). Tirgu Mures, Romania: IEEE, 2017: 1-6.