陈山杉, 董育宁. 结合特征选择的抗噪声网络流分类[J]. 智能计算机与应用,2024,14(4):238-243. DOI:10.20169/j.issn.2095-2163.240439

结合特征选择的抗噪声网络流分类

陈山杉, 董育宁
(南京邮电大学 通信与信息工程学院, 南京 210003)

摘 要: 在现代网络管理中,网络环境愈加复杂,网络流噪声已经成为不可忽视的因素,然而现有的抗噪声流分类方法在实际效果上仍有不尽人意之处。针对这一问题,本文提出了一种结合特征选择的抗噪声网络流分类方法 NNTC-FS,该方法采用投票机制判定噪声并搭建级联结构,实现先过滤再分类的线上任务。在公共数据集上的实验表明,NNTC-FS 能实现 90%以上的分类正确率,并在分类精度和时间性能上优于文献方法。

关键词: 网络流分类; 抗噪声; 投票; 特征选择; 级联模型
中图分类号: TP393 **文献标志码**: A **文章编号**: 2095-2163(2024)04-0238-06

Noise-resistant network traffic classification combined with feature selection

CHEN Shanshan, DONG Yuning
(School of Communication and Information Engineering, Nanjing University of Posts and Telecommunications, Nanjing 210003, China)

Abstract: Network flow classification plays an important role in modern network management. In the increasingly complex network environment, network flow noise has become a factor that cannot be ignored. However, the existing anti-noise network traffic classification methods are still unsatisfactory in practice. In view of this problem, this paper proposes a noise-resistant network traffic classification combined with feature selection method(NNTC-FS), which uses a voting mechanism to determine noise and constructs a cascade model to implement online tasks that filter before classifying. Experiments on public datasets show that NNTC-FS can achieve a classification accuracy of over 92%, and is superior to literature methods in accuracy and time performance.

Key words: network traffic classification; anti-noise; voting; feature selection; cascade model

0 引 言

在许多国家,合法拦截是一项任务。例如,中国禁止使用 Google 搜索引擎、禁止使用 Gmail 等[1]。网络流分类(Network Traffic Classification, NTC)为合法拦截提供了技术基础[2],保证了网络安全[3-4]。因此,近年来流分类技术越来越受到关注。

随着第五代移动通信技术与无线路由的普及,大数据时代带来便利的同时,NTC 也面临着新的挑战,复杂的网络中可能掺杂了噪声[5],降低了 NTC 的性能。为此,本文提出一种结合特征选择的抗噪声网络流分类方法(Noise-resistant Network Traffic Classification Combined with Feature Selection, NNTC-FS)。基于特征选择(Feature Selection, FS)的方法能提高分类的性能,减少分类时间。为了应对噪声,本文构建了投票分类器组,根据结果判定噪声;搭建了两个分类器的级联,实现先过滤再分类的线上分类。

1 相关工作

在网络流分类领域,研究者们主要关注基于机器学习(Machine Learning, ML)[6]和深度学习(Deep Learning, DL)[7]的方法。由于深度学习训练需要大量的时间和样本[8],而大数据时代,网络流分类也需要同时关注分类的速率[9],因此在线上网络流分类领域,传统机器学习更具优势。本节回顾了与本文相关的基于特征选择和抗噪声的传统机器学习流分类方法。

基金项目: 国家自然科学基金(61271233); 江苏省研究生科研与实践创新计划项目(KYCX21_0750)。
作者简介: 陈山杉(1998-),男,硕士研究生,主要研究方向:网络流识别与分类。
通讯作者: 董育宁(1955-),男,博士,教授,博士生导师,主要研究方向:多媒体通信,网络流识别与分类。Email:19900011@njupt.edu.cn
收稿日期: 2023-03-20

考虑到互相关性低的特征有利于分类，Zhang 等[10]提出 BoF(Bag of Flows)，用于衡量特征间的关系，并将包含 BoF 的 20 个统计特征作为 FS 的依据进行流分类。Perna 等[11]也考虑了相关性，还据此减少了特征的数目。实验证明，选取 10 个统计特征，分类准确率可以达到 95%。

新的特征同样也能提高分类精度。项等[12]提出了条件频率特征的定义与计算方法，表明新的特征能有效提高分类精度。Fahad 等[13]从 Moore 等[14]提出的特征列表中选择了一些特征，并引入贝叶斯核估计方法对流量进行分类。

针对流数据里的噪声，目前主流的两类方法是删除噪声样本和修改噪声样本标签为正常样本。为了删除数据集里的噪声，Wang 等提出了投票的解决方法，根据分类器组的投票结果按比例删除噪声样本，实验证明分类器精度最大提升 10%。除了针对样本的分布属性进行归类，在 FS 阶段，Wu 等[15]提出了一种新的基于一致性度量的特征选择和实例净化方法。由于删除噪声样本会降低结果的覆盖率，因此 Yuan 等利用多组分类器，对数据集进行分类，并根据多组结果为样本附加权值，为样本重新标记。Nicholson 等[16]提出了自训练校正和聚类两种方法，用于重新标记噪声样本。

但在实际流分类场景中，因为测试集噪声的存在，重标记方法可能难度较大。因此，本文选择删除噪声样本的方法。在线下训练阶段，设计了一种基于投票的噪声判定机制(Noise Decision based on Voting，NDV)，有效地选择出噪声；在线上分类阶段，级联分类器可以在线过滤噪声并进行分类。

2　方法

2.1　NNTC-FS 模型

本文方法 NNTC-FS 模型框架如图 1 所示，其中包括线下训练与线上分类 2 个阶段。在线下训练阶段，首先用 FS 方法筛选出合适的特征组合；其次，使用多个不同的分类器构建投票分类器组，利用 NDV 机制筛选出合适样本，划分出噪声训练集与正常样本训练集，用以训练噪声 - 非噪声二分类器 RF_2 和多分类器 RF_n。在线上分类阶段，利用 RF_2 - RF_n 的级联，实现先过滤再分类的流分类任务。

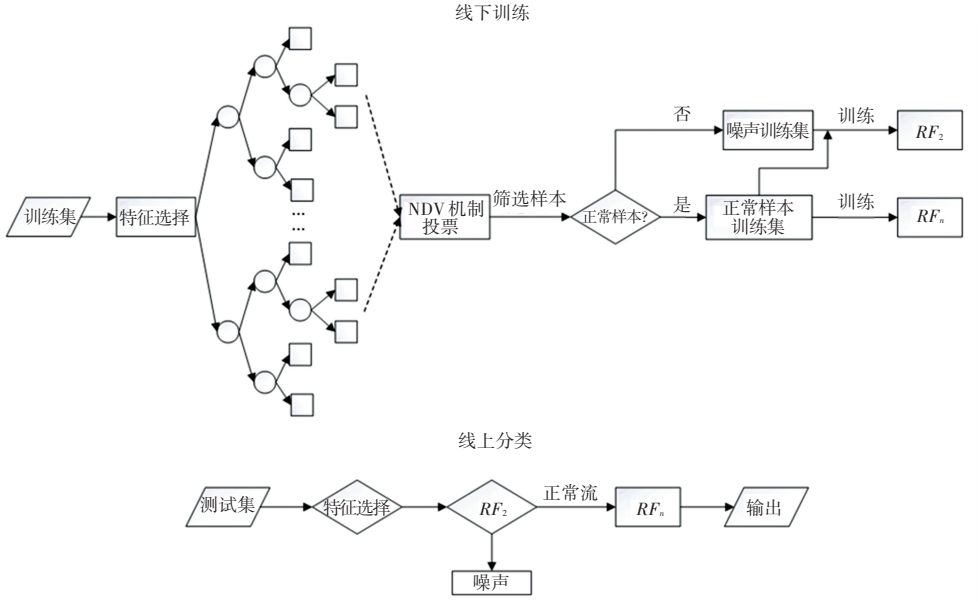


图 1　NNTC-FS 模型框架

Fig. 1　Framework of the proposed method CACV

2.2　特征选择

将采集的流数据划分为 1 s 的流段，且特征计算都是基于每条流的前 10 个数据包。本文共计算了 102 个统计特征与 29 个条件频率特征，并从 131 个特征中选出复杂度不超过 $O(n)$ 的特征。之后，再通过算法 1 对特征进行降维。

NNTC-FS 共进行了两次降维，目标是选择出数目少且作用大的特征子集。第一次降维采用过滤式方法，计算每个特征关于标签及特征之间的皮尔森相关系数[17]，从相关系数大于 0.9 的特征对中删除与标签相关性小的特征；第二次降维使用嵌入式方法，通过随机森林模型对剩下的特征进行排序，再逐个添加特征参与分类，根据分类结果寻找性能的拐点。在 ISCX 数据集[18]下本文方法 NNTC-FS 特征

数目与分类性能（*Acc*）的关系如图 2 所示。由此可见，当特征数目增加到 8 之后，*Acc* 增加缓慢，因此这 8 个特征为 ISCX 下分类的最优特征组。

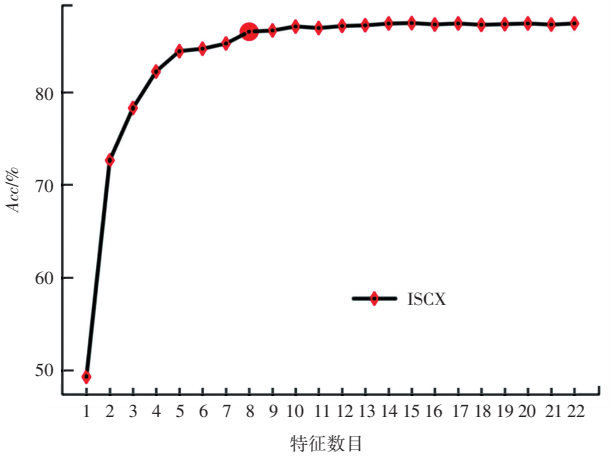


图 2 ISCX 下特征数目与 *Acc* 的关系图

Fig. 2 Relationship between the number of features and the *Acc* on ISCX

2.3 NDV 机制

考虑到标签噪声以及属性噪声的存在，本文设计了噪声判定准则 NDV 机制，同时考虑了投票结果与原样本标签的情况。具体流程以及判定准则见 NDV 算法：首先，是 5 个不同训练子集的 RFC 组成的投票分类器组；再根据不同的结果组合，将样本判定为噪声样本或正常样本。若预测结果间不符合准则，则为属性噪声，若结果与标签不符合准则，则为标签噪声，只有同时满足两个条件，才判定为正常样本（见 NDV 算法中 2-5 行）。

NDV 算法

输入 5 RF classifiers, *Nsam* samples,

Sample i owns 5 votes:

$V_i = \{v_{i1}, v_{i2}, v_{i3}, v_{i4}, v_{i5}\}$,

l label: y_i; $(y_i = 1,2,3,4,5)$

输出 The noise set *NoisD*, and normal data set *NormD*

For the samples V_i | $(i = 1,2\cdots Nsam)$

统计 V_i 中重复最多的个数 n_i，以及对应的值 v_i

if $(n_i \geqslant 3 \&\& v_i = y_i)$:

put sample i into *NormD*

else: put sample i into *NoisD*;

2.4 级联模型

线上分类模块不同于直接分类，NNTC-FS 采用级联模块（如图 3），先过滤再分类。用噪声样本与正常样本集训练 RF_2，同时可以通过改变两个训练集的比例，调节分类准确率与覆盖率，以满足不同的场景需求。

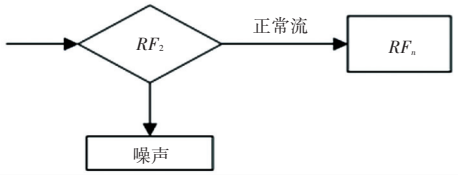


图 3 级联模块

Fig. 3 The cascade model

3 实验结果与分析

3.1 数据集与环境设置

由于 ISCX 数据集是一个认可度比较高的公共数据集，共含有 17 类数据样本（见表 1），因此实验选择 ISCX 进行，并通过方法的对比评价，比较了本文方法的分类和时间性能。实验采用 5 折交叉验证方法；训练集与测试集的样本比例为 4∶1。提取每条流的前 10 个数据包。

所有实验在硬件配置为 Intel i5 CPU@2.70 GHz，8 GB 内存的笔记本电脑完成，操作系统为 MacOS 64 位；用 Python 编程语言实现。在对比中，本文方法 NNTC-FS 级联模块选择的噪声样本与正常样本的比例为 1∶5，目的是保证覆盖率在 90%以上，以增加实验的可比性。

3.2 评价指标

评价包括对分类性能和计算性能的评估。分类性能可以量化成 5 种评价指标[19]，分别是总体准确率（Accuracy, *Acc*）、查准率（Precision, *P*）、查全率（Recall, *R*）、*F*1 测度（*F*1 - *score*）和覆盖率（Coverage）。其中，*Acc* 是指所有分类正确的样本占全部样本的比例；*P* 为预测是正例的结果中，正例的占例；*R* 是所有正例样本中被找出的比例；*F*1 测度是 *P* 和 *R* 的调和平均。计算性能主要分为特征提取的时间，另一方面是分类器的计算性能。计算公式如式（1）-式（5）所示。

$$Acc = \frac{TP + TN}{TP + FP + FN + TN} \tag{1}$$

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

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

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

$$Coverage = \frac{N_1}{N_0} \tag{5}$$

3.3　实验结果

3.3.1　分类性能比较

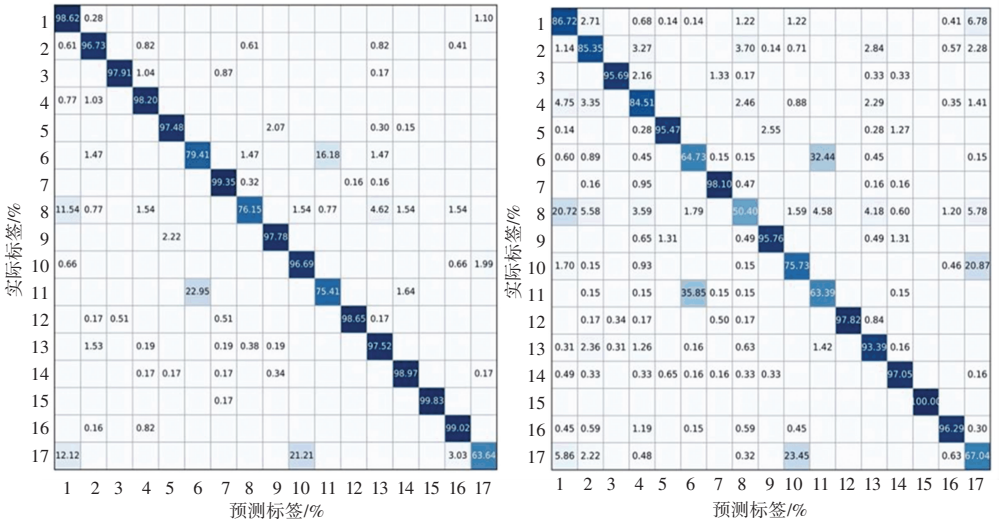
本文方法与 NSTC 方法[3]分类指标的比较结果见表 1。NSTC 方法的 *Acc* 为 88.56%，而 NNTC-FS 方法为 92.35%，相比提高 4 个点左右。

实验还比较了微观指标，逐类的 *F*1 测度与混淆矩阵（见表 1 和图 4）。由此可见，*F*1 测度越大，混淆矩阵对角线的方格颜色越深，则代表分类效果越好，无论从宏观或微观的角度来看，本文方法 NNTC-FS 的分类性能更优。

表 1　两种方法分类性能的对比

Table 1　Comparison of classification performance between the two methods

类别	*F*1 测度		类别	*F*1 测度		*Acc*	
	NNTC-FS	NSTC		NNTC-FS	NSTC	NNTC-FS	NSTC
Youtube	0.956 5	0.906 5	SFTP	0.727 3	0.557 3		
Vimeo	0.960 8	0.900 8	FTP	0.737 5	0.567 5		
Netflix	0.970 4	0.950 4	Facebook_audio	0.998 1	0.978 1		
Facebook_video	0.967 3	0.917 3	Hangsout_audio	0.960 4	0.910 4		
Hangsout_video	0.958 2	0.928 2	VoIPBuster	0.827 3	0.507 3	0.923 5	0.885 6
Skype_video	0.801 2	0.601 2	Bittorrent	0.988 2	0.988 2		
Facebook_chat	0.996 9	0.966 9	Email	0.991 2	0.971 2		
Hangsout_chat	0.768 6	0.608 6	Spotify	0.656 9	0.636 9		
Skype_file	0.959 7	0.979 7					



(a) NNTC-FS　　(b) NSTC

图 4　两种方法在 ISCX 上的混淆矩阵比较

Fig. 4　Comparison of confusion matrices between two methods on ISCX

3.3.2　时间性能比较

实验分别比较了每个阶段的消耗时间，结果见表 2。在特征提取模块，NSTC 所花时间高于本文方法，增加了大约两个数量级；在训练阶段，由于 NNTC-FS 借助了集成的思想，采用了多个分类器的级联，因此训练时间略长于 NSTC 方法。但综合线下与线上耗时，本文方法的时间性能表现更好。可见，NNTC-FS 方法在分类性能和时间性能上均优于 NSTC 方法。

表 2　在 ISCX 下不同方法时间性能对比

Table 2　Comparison of time performance of different methods on ISCX　　ms

时间	ISCX（17 类）	
	NSTC	NNTC-FS
特征提取时间	4.938 1	**0.019 5**
训练时间	**0.632 7**	0.974 1
识别时间	0.030 3	**0.008 8**
总时间	5.873 8	**1.002 4**

3.4 方法分析

本文方法 NNTC-FS 在 FS、噪声判定、与分类模型这 3 个部分做了改进,下面将详细分析和探究其原因。

3.4.1 特征选择对比

采用单个 RFC,基于 NNTC-FS 和 NSTC 的特征进行分类,分类结果以及特征的计算时间见表 3。由表中数据可见,本文方法选择的特征不仅数目少于 NSTC,耗时也更少。因为去除了冗余的复杂度高的特征,精度不会下降且计算时间缩短。同时,本文除了统计特征还增加了条件频率特征,这也可能会提升分类的精度。

表 3 两种方法基于特征的性能对比

Tabel 3 Performance comparison of two methods based features

参数	NNTC-FS	NSTC
特征数目	8	20
单特征(ms/样本)	0.002 3	0.020 5
总耗时(ms/样本)	0.018 4	0.410 0
Acc	0.892 5	0.866 7

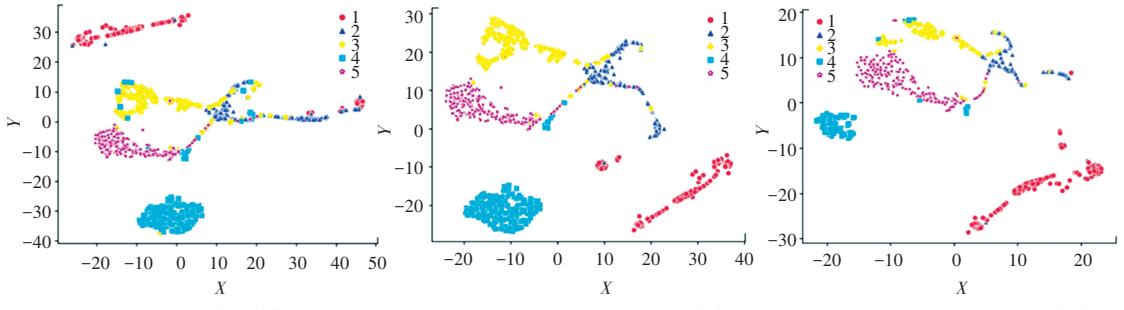
3.4.2 噪声判定对比

借助可视化工具 T-SNE 算法[20]将特征从多维降到二维,可以直观地比较噪声过滤前后各样本聚类的情况。随机选择的 5 类样本集和经过方法过滤的可视化结果如图 5 所示,其中 5 类样本用 5 种不同的颜色和形状表示。可以看到,有些类聚类较好,而有些样本中(如:2、3、5 类)有重合的样本点,即为难以区分的噪声点。

经过 NSTC 的过滤,样本间的噪声点减少,但样本点仍有不少重叠,且第 4 类正常样本也被删除;而遵循本文 NDV 机制处理后的样本集中,类间的重叠点减少,类间距离也会变大,聚类效果较好,可见 NDV 机制更有效。

3.4.3 分类模型

本文方法设置了 RF_2 与 RF_n 的级联模型(见图 3)。删除噪声能提高分类的精度,但代价是损失了一定的覆盖率。可以通过调整二分类器训练集噪声样本与正常样本的比例,改变测试阶段输出的结果,以灵活调节分类准确率与覆盖率。



(a) 原始样本集 (b) NNTC-FS过滤后样本集 (c) NSTC方法过滤后样本集

图 5 不同方法过滤后的样本集可视化图

Fig. 5 Visualization of sample sets filtered by different methods

本文方法 NNTC-FS 的具体表现如图 6 所示。RF_2 的噪声样本与正常样本训练集比例从 1∶0.3 变化到 1∶5。当比例最大时,过滤的噪声样本最多,*Acc* 最高且对应的 Coverage 最低;当比例最小时,恰恰相反。随着样本比例的调整,结果的准确率和覆盖率都有显著变化,但准确率始终能保持在较高的水平。因此,可以适应不同的应用场景。

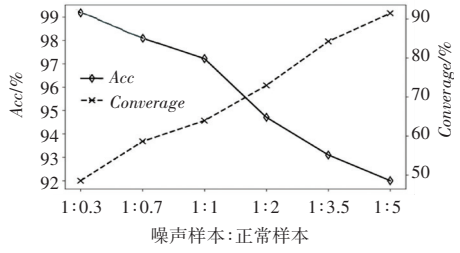


图 6 本文方法在 ISCX 上的 *Acc* 调节图

Fig. 6 *Acc* adjustment of CACV on ISCX

4 结束语

在流分类中,本文提出的 NNTC-FS 方法可对噪声进行在线识别与过滤。首先,基于 FS 本文能以较少的特征数目参与分类,优化了时间性能。其次,为了判定噪声,设计了 NDV 机制;为了过滤噪声,构造级联模块先实时过滤再完成分类。最后,可以通过调整二分类器训练集比例,实现结果分类准确率与覆盖率的灵活调节,以适用不同的应用场景。

在公共数据集 ISCX 上,本文方法 NNTC-FS 能实现较好的抗噪性与分类性能,用较快的分类速度,达到 90% 以上的分类精度。较文献方法 NSTC,NNTC-FS 在分类性能与时间性能上表现更好。

然而,本文方法仍存在一定的局限性,分类可能会为了实现较高的准确率而牺牲覆盖率。且含噪样

本中可能存在未知类。因此,在接下来的研究中,将考虑含噪样本中可能存在的未知类样本。

参考文献

[1] 唐玺博，张立民，钟兆根. 基于 ADASYN 与改进残差网络的入侵流量检测识别[J]. 系统工程与电子技术，2022，44(12)：3850-3862.

[2] 任家东，张亚飞，张炳,等. 基于特征选择的工业互联网入侵检测分类方法[J]. 计算机研究与发展，2022，59(5)：1148-1159.

[3] WANG B，ZHANG J，ZHANG Z L，et al. Noise-resistant statistical traffic classification[J]. IEEE Transactions on Big Data，2019，5(4)：454-466.

[4] YUAN W W，GUAN D H，MA T H，et al. Classification with class noises through probabilistic sampling [J]. Information Fusion，2017，41：57-67.

[5] QIAN H J，WEN Q S，SUN L，et al. RobustScaler：QoS-aware autoscaling for complex workloads [C]//Proceedings of 2022 IEEE 38th International Conference on Data Engineering (ICDE). Kuala Lumpur：IEEE，2022：2762-2775.

[6] SIMPSON K A，CZIVA R，PEZAROS D P. Seiðr：Dataplane assisted flow classification using ML [C]// Proceedings of 2020 IEEE Global Communications Conference. TaiWan：IEEE，2020：1-6.

[7] YANG L X，FINAMORE A，JUN F，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.

[8] TANG Pingping，DONG Yuning ，MAO Shiwen. Online traffic classification using granules [C]// Proceedings of IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS). Toronto：IEEE，2020：1135-1140.

[9] REZAEI S，LIU X. Deep learning for encrypted traffic classification：an overview[J]. IEEE Communications Magazine，2019，57(5)：76-81.

[10] ZHANG J，XIANG Y，WANG Y，et al. Network traffic classification using correlation information[J]. IEEE Transactions on Parallel and Distributed Systems，2012，24(1)：104-117.

[11] PERNA G，MARKUDOVAY D，TREVISANY M，et al. Online classification of RTC Traffic[C]// Proceedings of 2021 IEEE 18th Annual Consumer Communications and Networking Conference (CCNC). New York：IEEE，2021：1-6.

[12] 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.

[13] FAHAD A，TARI Z，KHALIL I，et al. Toward an efficient and scalable feature selection approach for internet traffic classification [J]. Computer Networks，2013，57(9)：2040-2057.

[14] MOORE A，ZUEV D，CROGAN M. Discriminators for Use in Flow-Based Classification [M]. Queen Mary and Westfield College，Department of Computer Science，2005.

[15] WU Z，DONG Yuning，WEI Hualiang，et al. Consistency measure based simultaneous feature selection and instance purification for multimedia traffic classification[J]. Computer Networks，2020，173：107190.

[16] NICHOLSON B，SHENG V S，JING Z，et al. Label noise correction methods[C]//Proceedings of 2015 IEEE International Conference on Data Science and Advanced Analytics (DSAA). Shanghai：IEEE，2015：1458-1462.

[17] EDELMANN D，MÓRI T F，SZÉKELY G J. On relationships between the pearson and the distance correlation coefficients[J]. Statistics and Probability Letters，2021，169(2021)：108960.

[18] LASHKARI A H，DRAPER-GIL G，MAMUN M，et al. Characterization of encrypted and VPN traffic using time-related features[C]// Proceedings of the International Conference on Information Systems Security and Privacy (ICISSP). Rome：IEEE，2016：407-414.

[19] 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.

[20] TAYLOR J，MERÉNYI E. Automating T-SNE parameterization with prototype-based learning of manifold connectivity [J]. Neurocomputing，2022，507(2022)：441-452.