计算机研究与发展
Journal of Computer Research and Development
DOI:10.7544/issn1000-1239.2020.20190691
57(12):2673-2682,2020

基于散度的网络流概念漂移分类方法

程 光[1,2,3] 钱德鑫[1,2,3] 郭建伟[4] 史海滨[1,2,3] 吴 桦[1,2,3] 赵玉宇[1,2,3]
[1](东南大学网络空间安全学院 南京 211189)
[2](教育部计算机网络和信息集成重点实验室(东南大学) 南京 211189)
[3](江苏省计算机网络技术重点实验室(东南大学) 南京 211189)
[4](华为技术有限公司西安研究所 西安 710075)
(gcheng@njnet.edu.cn)

A Classification Approach Based on Divergence for Network Traffic in Presence of Concept Drift

Cheng Guang[1,2,3], Qian Dexin[1,2,3], Guo Jianwei[4], Shi Haibin[1,2,3], Wu Hua[1,2,3], and Zhao Yuyu[1,2,3]
[1](*School of Cyber Science and Engineering, Southeast University, Nanjing* 211189)
[2](*Key Laboratory of Computer Network and Information Integration (Southeast University), Ministry of Education, Nanjing* 211189)
[3](*Jiangsu Provincial Key Laboratory of Computer Network Technology (Southeast University), Nanjing* 211189)
[4](*Xi'an Research Institute, Huawei Technologies Co., Ltd., Xi'an* 710075)

Abstract Due to the high dynamic variability, suddenness and irreversibility of network traffic, the statistical characteristics and distribution of traffic may change dynamically, resulting in a concept drift problem based on the flow-based machine learning method. The problem of concept drift makes the classification model based on the original data set worse on the new sample, which causes the classification accuracy to decrease. Based on this, a classification approach based on divergence for network traffic in presence of concept drift, named ECDD (ensemble classification based on divergence detection) is proposed. The method uses a double-layer window mechanism to track the concept drift. From the perspective of information entropy, the Jensen-Shannon divergence is used to measure the difference of data distribution between old and new windows, so as to effectively detect the concept drift. This paper draws on the idea of incremental ensemble learning, trains a new classifier on the concept drift traffic based on the pre-retention classifier, and replaces the classifier with the original performance degradation according to the classifier weight, so that the ensemble classifier is effectively updated. For common network application traffic, this paper constructs a concept drift data set according to different application feature distributions. This paper compares the method with common concept drift detection methods and the experimental results show that the method can effectively detect concept drift and update the classifier, showing better classification performance.

收稿日期:2019-09-25;**修回日期**:2020-04-13
基金项目:国家重点研发计划项目(2018YFB1800602,2017YFB0801703);教育部-中国移动科研基金项目(MCM20180506);国家自然科学基金项目(61602114);赛尔网络下一代互联网技术创新项目(NGIICS20190101,NGII20170406)
This work was supported by the National Key Research and Development Program of China (2018YFB1800602, 2017YFB0801703), the Ministry of Education-China Mobile Research Fund Project (MCM20180506), the National Natural Science Foundation of China (61602114), and the CERNET Innovation Project (NGIICS20190101, NGII20170406).
通信作者:钱德鑫(dxqian@njnet.edu.cn)

Key words concept drift; machine learning; Jensen-Shannon divergence; incremental ensemble learning; traffic classification

摘 要 网络流量特征分布的动态变化产生概念漂移问题,造成基于机器学习的网络流量分类模型精度下降.定期更新分类模型耗时且无法保证分类模型的泛化能力.基于此,提出一种基于散度的网络流概念漂移分类方法(ensemble classification based on divergence detection, ECDD),采用双层窗口机制,从信息熵的角度出发,根据流量特征分布的JS散度,记为 JSD(Jensen-Shannon divergence)来度量滑动窗口内数据分布的差异,从而检测概念漂移.借鉴增量集成学习的思想,检测到漂移时对于新样本重新训练出新的分类器,之后通过分类器权值排序,保留性能较高的分类器,加权集成分类结果对样本进行分类.抓取常见的网络应用流量,根据应用特征分布的不同构建概念漂移数据集,将该方法与常见的概念漂移检测方法进行实验对比,实验结果表明:该方法可以有效地检测概念漂移和更新分类器,表现出较好的分类性能.

关键词 概念漂移;机器学习;JS散度;增量集成学习;流量分类

中图法分类号 TP393

随着互联网的飞速发展,网络新应用、新协议层出不穷,网络应用程序越来越复杂;同时网络中的加密流量比例随着人们对于隐私的重视而不断扩大.这些问题使得传统的流量分类方法面临巨大的挑战,逐渐从基于端口和深度包检测技术(deep packet inspection, DPI)[1]的流量分类方法转向基于机器学习[2-3]的流量分类方法.基于机器学习的流量分类比较依赖于训练样本的环境.由于网络流量具有高度动态变化性、突发性、不可逆性等特点,因此流量的统计特征和分布会发生动态变化,导致基于流特征的机器学习方法产生概念漂移问题[4-5].网络流量概念漂移问题使得建立在原始数据集上的分类模型在新样本上适用性变差,造成分类器分类精度下降.

针对概念漂移问题国内外学者进行了大量的研究,并取得众多成果.然而这些方法还存在一些不足:1)当前大多数解决概念漂移的方法都是基于分类错误率[6]来检测概念漂移,带来的问题是获取样本标签需要耗费大量时间和资源,同时基于分类错误率的检测方法无法应对类不平衡[7]的网络流.2)当检测到概念漂移时,若只在新流量上训练分类器会导致历史知识的丢失[8],若将不同时期获得的所有流量重新训练分类器则会带来较大的性能问题[9].一个理想的网络流分类模型应该能增量地学习并适应多种网络流量类型的变化,因此设计出能应对概念漂移的网络流量分类方法具有重要的意义.

针对上述问题,本文提出了一种基于散度的网络流概念漂移分类方法(ensemble classification based on divergence detection, ECDD).该方法首先采用双层窗口机制跟踪概念漂移,借鉴信息论的思想,通过JS散度来度量新旧窗口之间数据分布的差异,从而有效地检测概念漂移.本文采用增量集成学习的方法,当检测到概念漂移时采用新样本重新训练分类器,接着通过分类器权值比较的方式保留性能较好的分类器,从而实现对于集成分类模型的有效更新.

1 相关研究

近年来,许多学者对于概念漂移问题提出了大量的算法与模型,目前概念漂移检测方法大体上分为2类:1)依据概念漂移造成的后果进行检测,例如依据分类准确率下降来检测概念漂移;2)依据概念漂移产生的原因进行检测,例如依据数据分布是否发生变化来检测概念漂移.

对于使用分类错误率来检测概念漂移,相关研究为:Gama等人[10]提出根据当前分类器的分类错误率来检测概念漂移的DDM算法,该方法持续监测分类器的分类错误率,对于突变式的概念漂移表现优越,但不能有效地检测到渐变式的概念漂移;Nishida等人[11]提出了基于伯努利分布检测概念漂移的算法EDDM,该方法同时依据模型的错误率变化与错误率之间的距离判断概念漂移,不仅对突变性的概念漂移检测良好,同时也提高了对渐变概念漂移的检测效果;Bifet等人[12]提出了自适应滑动窗口算法ADWIN,该算法用可变的滑动窗口来检测随时间变化的数据流,通过比较不同子窗口之间的分类错误率均值差异来判断是否发生概念漂移;

Nishida 等人[13]提出一种突变式概念漂移检测方法 LID,依据当前模型的信息度和分类精度来检测概念漂移.

上述方法均基于分类准确率检测概念漂移,标记样本需要耗费大量时间和资源,因此很多学者基于概念漂移产生的原因,即数据分布的变化,取得了相应的研究成果:Borchani 等人[14]基于半监督学习的方法提出基于距离的漂移检测,该方法使用滑动窗口将数据流划分为不同的子集,基于 KL 散度(Kullback-Leibler divergence, KLD)计算差异性来检测概念漂移,该方法使用的 KL 散度取值范围较大,并存在距离不对称等问题;Alippi 等人[15]利用中心极限定理,设计了不依赖数据分布模型、不需要任何先验信息的概念漂移检测算法;Dries 等人[16]基于统计学方法,使用对不同概念漂移类型设定不同阈值来检测漂移,克服了单一标准检测概念漂移的弊端,提高了对不同类型概念漂移的检测准确率;潘吴斌等人[17]基于概念漂移发生时属性信息熵会发生变化来检测概念漂移,该方法对于不同类型的概念漂移均可以有效检测并更新分类器,使得分类器具有良好的泛化性能.

由于网络流量的动态性,例如不同时间、地点、网络应用更新、新型网络应用出现等情况,使得网络业务分布发生变化,即会产生概念漂移问题,概念漂移使得分类模型的分类准确性下降,影响分类器的泛化性能.

将网络流量统计特征表示为 $\boldsymbol{A}=(a_1,a_2,\cdots,a_n)$,用 B 来表示网络流量所属类别,根据贝叶斯公式可得,分类器可以定义为期望函数 $f:\boldsymbol{A}\rightarrow B$:

$$f(\boldsymbol{A})=\arg\max_{b\in B}P(b\mid\boldsymbol{A})=\arg\max_{b\in B}\frac{1}{P(\boldsymbol{A})}P(b)\prod_{i=0}^{n}P(a_i\mid b),\tag{1}$$

由式(1)可得,流统计特征 a_i 的变化会影响类条件概率密度 $P(a_i/b)$,从而影响到 $P(b/\boldsymbol{A})$;同时类别先验概率 $P(b)$的变化也会影响到 $P(b/\boldsymbol{A})$,从而影响分类结果.

本文提出了一种基于散度的概念漂移流量分类方法 ECDD,该方法基于双层窗口检测机制,依据特征属性的 JS 散度差异来判断数据分布的变化,从而检测概念漂移;检测到概念漂移时需要更新分类器,本文采用增量集成学习的方式引入漂移流量更新分类器,在保留之前分类器的基础上剔除性能下降的分类器,从而使得分类器得到有效地更新,有效地应对网络流量概念漂移问题.

2　基于散度的概念漂移分类方法

本节将首先介绍基于散度的网络流概念漂移检测算法,然后描述在概念漂移点采用增量集成学习的方式更新分类器的方法.

2.1　基于散度的概念漂移检测算法

概念漂移问题使得分类模型对新流量的适用性变差,难以保持稳定的识别准确性.仅依靠经验定期更新分类器会浪费大量的时间和人力,及时准确地检测到概念漂移,从而对分类器进行更新至关重要.基于分类错误率的概念漂移检测算法需要获得检测样本的真实标签,浪费大量人力资源,同时不能很好地适用于类不平衡的网络流.

本文提出了一种基于 JS 散度的概念漂移检测算法 ECDD.在信息论中,相对熵(relative entropy)又称为 KL 散度,是衡量 2 个概率分布之间的差异性的指标,假设用 X 表示某随机变量,该随机变量上的 2 个概率分布为 $P(x)$和 $Q(x)$,则它们的 $KLD(P\parallel Q)$定义为

$$KLD(P\parallel Q)=\sum_{x\in X}P(x)\ \mathrm{lb}\ \frac{P(x)}{Q(x)},\tag{2}$$

可以得出,KLD 的取值范围为$[0,+\infty]$,当 2 个随机分布相同时,它们的相对熵为 0,当 2 个随机分布的差别增大时,它们的相对熵也会增大.由于 KLD 取值范围为$[0,+\infty]$,对于 2 组随机分布数据,无法准确地用 KLD 来度量 2 组数据分布的差异,对相似度的判别较为粗略.

尽管 KLD 从直观上是一个度量或距离函数,但它并不是一个真正的度量或距离,因为它不具有对称性,即:

$$KLD(P\parallel Q)\neq KLD(Q\parallel P),\tag{3}$$

KLD 的非对称性对求解 2 组数据分布是否相似存在一定偏差.

JS 散度是 KL 散度的一种变形,它解决了 KL 散度的不对称问题,对相似度的判别更确切,对于 2 个概率分布 $P(x)$和 $Q(x)$,$JSD(P\parallel Q)$的计算公式为

$$JSD(P\parallel Q)=\frac{1}{2}KLD\left(P\parallel\frac{P+Q}{2}\right)+\frac{1}{2}KLD\left(Q\parallel\frac{P+Q}{2}\right).\tag{4}$$

由式(4)可得 JSD 是 2 个 KLD 的累加,将式(4)中 KLD 展开得:

$$JSD(P \| Q)=\sum_{x\in X}(P(x)\ \mathrm{lb}\ \frac{2P(x)}{P(x)+Q(x)}+ Q(x)\ \mathrm{lb}\ \frac{2Q(x)}{P(x)+Q(x)}), \tag{5}$$

JSD 的取值范围为[0,1],即对于 2 组随机分布变量,分布相同则是 0,分布完全不同则为 1.相较于 KLD,JSD 对相似性度量的判别更加确切.同时,JSD 解决了 KLD 的非对称问题,即 JSD 满足对称性:

$$JSD(P \| Q)=JSD(Q \| P), \tag{6}$$

JSD 的对称性使得在计算 2 组数据分布间差异时不会产生偏差.

基于 JS 散度的取值范围更准确以及其满足对称性,本文引入 JS 散度来度量特征分布的变化,从而检测概念漂移.

本文采用双层滑动窗口的方式,在概念漂移检测流程中一直维护 2 个滑动窗口,分别表示较旧概念的数据和新概念的数据.考虑到由于类不平衡会使得特征分布发生变化,从而对概念漂移检测结果产生影响,本文先对 2 个窗口内样本进行分类,依据分类结果计算 2 个窗口中样本特征分布之间的 JS 散度,根据 JS 散度判断概率分布是否发生变化,从而检测是否发生了概念漂移.

定义时刻 t_i,用 $Win_{\mathrm{old},i}$ 表示最近一次检测到概念漂移时的窗口,$Win_{\mathrm{new},i}$ 代表当前检测窗口,包含最新的样本实例.用 $JSD(Win_{\mathrm{old},i} \| Win_{\mathrm{new},i})$ 表示 2 个窗口内数据分布的距离,定义漂移检测阈值 τ,若存在某特征对应的 2 组数据分布的 $JSD(Win_{\mathrm{old},i} \| Win_{\mathrm{new},i})>\tau$,则判断发生概念漂移,称 t_i 为概念漂移点.当新样本加入时,窗口 $Win_{\mathrm{new},i}$ 向前滑动,每次更新,检测是否存在 $JSD(Win_{\mathrm{old},i} \| Win_{\mathrm{new},i})>\tau$,如果是,报告发生概念漂移,重复整个过程.算法 1 描述了概念漂移检测算法的流程:行①~⑦使用分类器对滑动窗口内样本进行分类,获得样本特征向量和分类结果;行⑧到结尾计算 2 个窗口内每类样本对应特征的 JS 散度,与漂移检测阈值 τ 比较,若大于阈值,则判断发生概念漂移,否则窗口向前滑动,重复整个流程直到数据流结尾.

算法 1. 概念漂移检测算法.

输入:分类器 C、滑动窗口大小 δ、特征向量 $\boldsymbol{R}^d$、漂移检测阈值 τ;

输出:概念漂移检测结果.

① $Win_{\mathrm{old},i}$ 表示在时刻 t_i 的 δ 个样本;

② $Win_{\mathrm{new},i}$ 表示在 $Win_{\mathrm{old},i}$ 后的 δ 个样本;

③ 当滑动窗口未到达数据流尾部时

④ 用 C 对 $Win_{\mathrm{old},i}$,$Win_{\mathrm{new},i}$ 样本分类;

⑤ $S_{\mathrm{old},i}$ 表示 $Win_{\mathrm{old},i}$ 的分类结果;

⑥ $S_{\mathrm{new},i}$ 表示 $Win_{\mathrm{new},i}$ 的分类结果;

⑦ 计算共有的类别 $y_{\mathrm{intersection},i}$;

⑧ 遍历 $y_{\mathrm{intersection},i}$

⑨ 遍历特征向量 $\boldsymbol{R}^d$

⑩ 若存在 2 组样本的 JS 散度大于 τ

⑪ 报告在时刻 t_i 发生概念漂移;

⑫ 清空滑动窗口,回到步骤①;

⑬ 将新窗口 $Win_{\mathrm{new},i}$ 滑动 $SlideSize$ 步长;

⑭ 概念漂移检测结束.

2.2 增量集成分类算法

本文借鉴集成学习里 bagging[18] 的思想构建集成学习初始分类器:假设集成分类器要构建 k 个基分类器,给定包含 m 个样本的原始训练样本集,通过 k 次随机采样,得到 k 个大小均为 m 的采样集 $S_1,S_2,\cdots,S_k$,对于这 k 个采样集,训练出 k 个基分类器,使用集成策略将这 k 个基分类器结合起来,形成最终的集成分类器.

借鉴自助采样法(bootstrap sampling)[19] 来进行随机采样:给定包含 m 个样本的数据集,有放回地进行 m 次样本的随机抽取,经过 m 次随机采样操作,将得到含 m 个样本的采样集,初始训练集中有的样本在采样集里多次出现,有的则从未出现.重复上述操作 k 次,这样可以得到 k 个相互独立的样本集,从而使得集成中的个体学习器尽可能相互独立,获得泛化性能强的集成.

选取 C4.5 决策树作为集成学习的基分类器,C4.5 决策树的分类模型易于理解,且具有较好的分类效果和效率[20-21].同时,C4.5 决策树在模型构建过程中不依赖于网络流分布,具有较好的分类稳定性.

考虑到基分类器性能不同对于其分类结果应持有不同的信心,因此本文根据基分类器的分类性能赋予其权重.用 $C_i(1\leqslant i\leqslant k)$ 表示基分类器,$W_i(1\leqslant i\leqslant k)$ 表示分类器 C_i 的权重,权重 W_i 与分类器 C_i 在其训练集 S_i 上的分类准确率成正比,定义 Acc_i 为基分类器 C_i 在其训练集 S_i 上的分类准确率,则该分类器对应的权重 W_i 计算为

$$W_i=\alpha\times Acc_i, \tag{7}$$

其中,$\alpha>0$ 为正相关系数.算法 2 描述了集成学习分类器的构造方法.行①~⑧描述了初始集成分类器构造的整体流程,行②~④为获取每个基分类器的训练样本集并训练基分类器,行⑤~⑦为计算基分类器权重并将基分类器加入到集成分类器 C 中.

算法 2. 集成学习初始分类器构造算法.

输入:基分类器数目 k、初始训练样本集 S;

输出:k 个带权重的基分类器集合 C.

① 遍历基分类器集合

② 从 S 中随机采样得到训练集 S_i;

③ 根据 S_i 训练基分类器 C_i;

④ 计算 C_i 的分类准确率 Acc_i;

⑤ 根据 Acc_i 计算 C_i 的权重 W_i;

⑥ 将 C_i 加入到集成学习分类器中;

⑦ 完成加权集成分类器的构建;

⑧ 返回加权集成分类器.

当检测样本到来时,集成分类器对每个基分类器的分类结果加权求和,从而得到检测样本的最终分类结果.对于网络流量分类任务,定义待检测样本为 x,分类器 C_i 将从类别标记集合 $label\_set=\{1,2,\cdots,N\}$ 中预测出一个标记作为样本 x 的分类结果,用 $C_i^j(x)$ 表示分类器 C_i 对 x 的分类结果为 j $(1\leqslant j\leqslant N)$,则 x 的最终分类结果 $H(x)$ 为

$$H(x)=\arg\max_j\sum_{i=1}^{k}W_iC_i^j(x). \tag{8}$$

本文采用增量集成学习的方法来进行分类器的更新,从而在检测到概念漂移时有效地更新分类器.检测到概念漂移时,对于新样本重新训练出新的分类器,在保留原先集成分类器的基础上,通过权值比较选取性能最优的分类器,从而提高集成分类器的泛化能力.

定义 S_{new} 为检测到的概念漂移样本集合,在 S_{new} 上学习出一个新的分类器 C_{new},分类器 C_{new} 的权重 W_{new} 采用与式(7)相同的方法获得.对于每个基分类器 C_i $(1\leqslant i\leqslant k)$,根据其在 S_{new} 上的分类准确率更新各自的权重.对于 C_{new} 和基分类器 C_i $(1\leqslant i\leqslant k)$,按照分类器权重排序选取前 k 个分类器构成新的集成分类器.算法 3 描述了增量集成学习算法流程.

算法 3. 增量集成学习算法.

输入:k 个预先训练的基分类器集合 C、概念漂移样本集 S;

输出:更新权重的 k 个基分类器集合 C.

① 在 S 上训练出新的分类器 C_{new};

② 计算 C_{new} 的权重 W_{new};

③ 遍历预先的基分类器集合 C

④ 计算基分类器 C_i 的分类准确率 Acc_i;

⑤ 根据分类准确率计算各自的权重 W_i;

⑥ 根据权重排序选取前 k 个分类器;

⑦ 返回更新后的加权集成分类器.

3　实验与分析

3.1　实验数据集

网络应用层出不穷,相同网络业务下的不同应用可能会存在特征方面的差异,即产生概念漂移问题,这对网络业务识别带来极大的挑战,造成网络业务识别准确率下降.

本文对常见的 8 种网络业务:视频(video)、音频(music)、网页浏览(Web)、发送接收图片语音

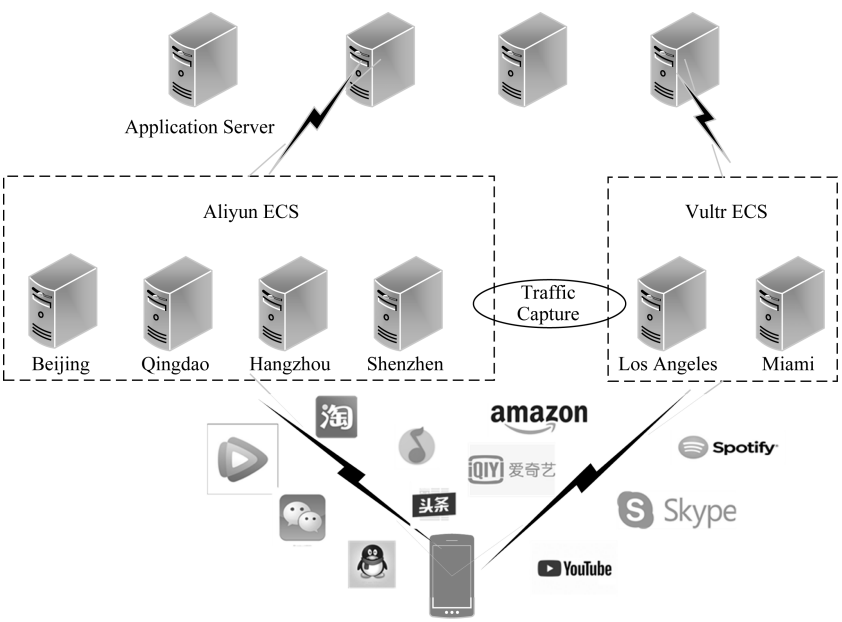


Fig. 1　Capturing traffic at different locations

图 1　不同地点采集流量

(picture&voice)、上传下载大文件(file)、文本聊天(chat)、视频通话(video call over IP)、语音通话(voice call over IP)进行流量抓取,每种网络业务根据该业务下应用的下载量,选择 3 种不同的应用分别进行流量采集.

考虑到终端抓取的流量可能无法完全反映网络的真实情况,本文使用 tcpdump 在 VPN 服务器上进行流量的抓取,抓取到的流量更远离终端,也更能实际反映出业务过程中的网络情况.

同时,考虑到 CDN 对流量特征的影响,采集流量时通过部署不同地点的服务器共同采集,从而尽可能地使采集到每种应用的流量能较真实地反应该应用整体的流量特征.

样本采集示意图如图 1 所示,分别选取不同服务器地址进行流量采集,从而使采集到的样本较好地反映该类网络应用的特征情况.

样本集具体构成情况如表 1 所示,对于每类网络业务,各进行 300 次独立拨测,根据组流后的结果构造出 300 条网络流量样本.其中每类网络业务由 3 种网络应用组成,每种网络应用各独立拨测 100 次,从而对于每类网络应用均获得互相独立的 100 条网络流量样本.

Table 1 Experimental Data Set Composition

表 1 实验数据集组成情况

Services	Applications	Number of Samples
Video	YouTube/QQLive/IQIYI	100/100/100
Music	Spotify/QQMusic/KuGou	100/100/100
Web	Amazon/Taobao/TouTiao	100/100/100
Picture&Voice	Skype/WeChat/QQ	100/100/100
File	Skype/WeChat/QQ	100/100/100
Chat	Skype/WeChat /QQ	100/100/100
Video Call over IP	Skype/WeChat /QQ	100/100/100
Voice Call over IP	Skype/WeChat /QQ	100/100/100

本文实验平台为 Windows 10,CPU 为酷睿 i5,内存为 12 GB,开发环境为 python 3.0.

3.2 特征选择

对于如表 1 所示数据经过组流、流统计特征处理后,每条样本均提取出 81 维特征.为了提高分类模型的分类性能和简化分类模型,需要进行特征选择,删除冗余、无效的特征.例如实际接收到的字节数等特征针对某些网络业务可能区分性不足,而平均码率、平均包大小等对不同的网络业务可能区分较明显.

本文结合常见的特征选择方法,充分考虑特征间的相关性、信息增益、信息增益率,综合各特征选择算法(InfoGain,GainRatio 等)选取最优特征,最终选择出了 11 维特征,如表 2 所示:

Table 2 Feature Selection Result

表 2 特征选择结果

Feature Index	Feature	Description
0	trans_protocol	Transport layer protocol type
1	duration	Duration of the flow in microsecond
2	pkt_len_mean	Mean size of packet
3	fwd_ pkt_len_mean	Mean size of packet in forward
4	bwd_ pkt_len_mean	Mean size of packet in backward
5	flow_rate	The average rate
6	fwd_flow_rate	The average rate in forward
7	bwd_flow_rate	The average rate in backward
8	pke_num_mean	The average number of packets
9	fwd_pkt_num_mean	The average number of packets in forward
10	bwd_pkt_num_mean	The average number of packets in backward

3.3 评估指标

本文采用基于散度的方法解决流量识别中的概念漂移问题,在使用机器学习算法进行流量识别的过程中,准确率是算法性能评估常用的指标,真正例 TP 表示将正类正确预测为正类数,假正例 FP 表示将负类错误预测为正类数,假负例 FN 表示将正类错误预测为负类数,真负例 TN 表示将负类正确预测为负类数.根据这 4 个指标,可以得到准确率($Accuracy$)、精确率($Precision$)、召回率($Recall$)和综合评价($F1\_score$)的计算方法:

$$Accuracy=\frac{TP+TN}{TP+FP+TN+FN}, \tag{9}$$

$$Precision=\frac{TP}{TP+FP}, \tag{10}$$

$$Recall=\frac{TP}{TP+FN}, \tag{11}$$

$$F1\_score=\frac{2\times Precision\times Recall}{Precision+Recall}. \tag{12}$$

同时,为了评估概念漂移检测方法的准确性及有效性,本文采用误报率(FPR)和漏报率(FNR)来评估.误报率和漏报率越高,表示概念漂移检测性能越差,误报率、漏报率分别计算为

$$FPR=\frac{FP}{TN+FP}, \tag{13}$$

$$FNR=\frac{FN}{TP+FN}. \tag{14}$$

3.4　概念漂移检测结果

3.4.1　概念漂移样本集构造

对于网络业务识别,由于同类网络业务下可能会包含不同的网络应用,各应用间可能会存在特征分布上的差异,从而使得该业务产生概念漂移问题,导致对于网络业务的识别准确率下降.

对表 1 所示的网络流量进行特征分析,寻找同类业务不同应用之间是否存在特征分布的差异.如图 2 所示为视频业务中,爱奇艺视频 IQIYI 与腾讯视频 QQLive 在平均速率特征上的分布直方图.

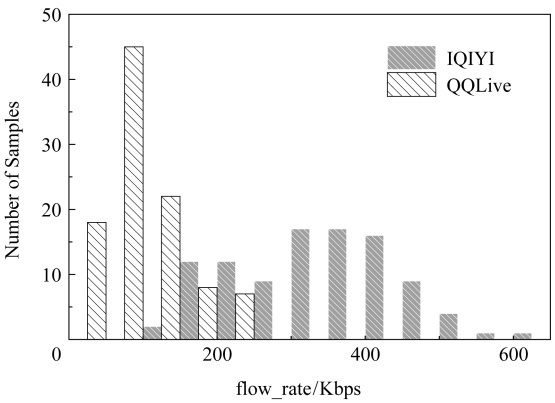


Fig. 2　Video app distribution differences in flow_rate

图 2　视频应用在平均速率上的分布差异

由图 2 可以看出爱奇艺视频与腾讯视频在平均速率这一特征上的分布差异.基于此方法,寻找表 1 所示网络业务不同应用间存在的特征分布差异,根据应用间特征的差异构造概念漂移数据集,表 3 所示为构造出的 5 种概念漂移数据集.

Table 3　Concept Drift Data Set

表 3　概念漂移数据集构成

Sample Set	Services	Applications	Number of Samples
Concept_1	Initial Samples	Initial Samples	600
Concept_2	Video	QQLive,IQIYI	600
Concept_3	Chat	WeChat,QQ	600
Concept_4	Music	KuGou,QQMusic	600
Concept_5	Video	QQLive,YouTube	600

如表 3 所示,通过更换某类网络业务下的应用构造出不同的概念漂移数据集,为了模拟真实网络环境中样本的情况,上述每个概念中的样本均随机打乱.

3.4.2　基于散度的概念漂移检测结果

本文依据 2 个窗口内数据的分类结果,使用 JS 散度度量特征间数据分布的差异,从而检测概念漂移.为了验证基于 JS 散度的概念漂移检测算法的有效性,本文做了 2 组对比实验.

以 Concept_1 样本集为例,将 Concept_1 样本集划分为大小相同的 2 个样本集,分别为 Concept_1_a 与 Concept_1_b,2 个样本集来自于同一个概念,依据分类结果对每个特征计算 JS 散度,结果如图 3 所示:

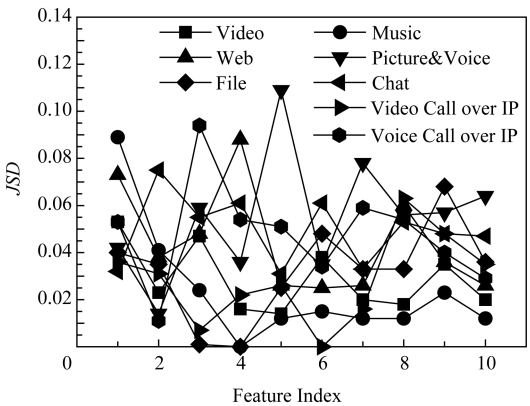


Fig. 3　JS divergence results for the same concept data set

图 3　相同概念数据集的 JS 散度计算结果

如图 3 所示,相同概念的样本集的特征计算出的 JS 散度值均较小(均小于 0.2),表示 2 个样本集特征分布差异较小,即可以判断没有发生概念漂移.

对比实验为 Concept_1 数据集和 Concept_2 数据集,由表 3 可以得出 Concept_1 与 Concept_2 数据集视频类业务存在特征差异,对这 2 个不同概念的数据进行 JS 散度计算,结果如图 4 所示:

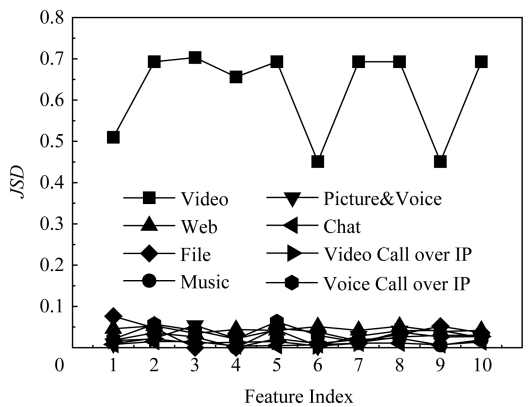


Fig. 4　JS divergence results for different concept data sets

图 4　不同概念数据集 JS 散度计算结果

如图 4 所示,Video 样本计算出的 JS 散度值较大

(图 4 中矩形标志的折线),表明 Video 类业务特征分布的差异较大,即可能发生了概念漂移.

上述实验窗口大小均取待检测数据集大小,并且没有定义漂移检测阈值,实验的目的是依据 JS 散度的检测结果与原始数据本身是否含有概念漂移进行比对,从而验证基于 JS 散度检测概念漂移的有效性.

3.4.3 检测窗口大小与检测阈值

本文采用滑动窗口的方式检测概念漂移,滑动窗口过大可能会造成概念漂移检测时漏报率增加,窗口过小可能会造成概念漂移检测时误报率增加,同时本文根据特征分布的 JS 散度衡量特征分布之间是否存在明显差异,从而判断是否发生概念漂移,JS 散度的阈值大小也影响着检测算法的性能.

为了确定合适的窗口大小与检测阈值,采用误报率与漏报率来评估不同窗口大小与阈值的检测性能,设置窗口大小为 100～250,阈值为 0.4～0.6,在不同的窗口与阈值情况下测试误报率与漏报率,结果如表 4 所示.

如表 4 所示,当检测阈值不变时,随着检测窗口变大,误报率呈现降低的趋势,漏报率呈现增加的趋势;同时当检测窗口大小不变时,随着检测阈值增大时,误报率也呈现降低趋势,漏报率呈现增加趋势.

综合误报率与漏报率的综合表现,选择窗口大小为 150,检测阈值为 0.5,从而实现较低的误报率与漏报率.

3.4.4 检测方法比较

DDM 和 STEPT 是 2 种常见的概念漂移检测方法[10].这里将这 2 种常见的概念漂移检测方法与 ECDD 方法进行对比,测试各算法的概念漂移检测性能.

Table 4 Detection Window and Threshold

表 4 检测窗口与检测阈值

Window	Threshold	Detections	*FPR*	*FNR*
100	0.4	18	0.778	0
	0.5	10	0.6	0
	0.6	7	0.571	0.143
150	0.4	6	0.333	0
	0.5	4	0	0
	0.6	5	0	0.25
200	0.4	4	0.25	0.25
	0.5	4	0.25	0.25
	0.6	5	0.4	0.2
250	0.4	4	0.25	0.25
	0.5	4	0.25	0.25
	0.6	4	0.25	0.25

DDM 方法根据分类错误率检测概念漂移,根据默认配置设置其警告级别为 2.0,漂移级别为 3.0;STEPT 方法采用统计检验来检测概念漂移,根据默认配置设置检测窗口大小为 30,警告显著性水平为 0.05,漂移显著性水平为 0.03.ECDD 方法则使用上述实验确定的滑动窗口大小为 150,漂移检测阈值为 0.5.

如表 5 所示为 ECDD 方法与 DDM,STEPT 方法在检测概念漂移的实验对比结果.

Table 5 Comparison of Detection Methods

表 5 检测方法比较

Methods	*FPR*	*FNR*
ECDD	0	0
DDM	0	0.33
STEPT	0.5	0.17

ECDD 方法相比较 DDM 和 STEPT 方法可以较有效地检测到概念漂移,综合误报率和漏报率,ECDD 在概念漂移检测性能上优于 DDM 和 STEPT 方法.

3.5 网络流分类结果

3.5.1 基分类器数目选择

对于给定的初始训练集,基分类器数目可能会对分类准确率产生影响.本文采用 C4.5 作为基分类器,以 Concept_1 为测试集,图 5 描述了在不同的基分类器数目下集成学习模型分类性能的变化情况.

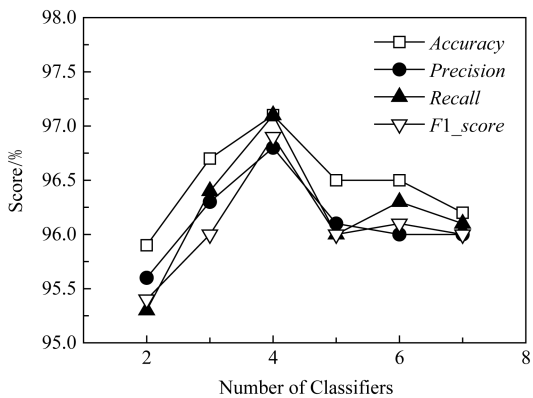


Fig. 5 The effect of the number of base classifiers

图 5 基分类器数目对分类效果影响

由图 5 可见,当分类器数目为 4 时,分类器的准确率、精确率、召回率和综合评价均达到最高,分类效果最优.减小或增加分类器数目时,各项评价指标都会出现相应的下降,因此本文集成学习基分类器数目选择 4,使分类模型性能最优.

3.5.2　分类准确率

本文根据准确率和综合评价评估算法的分类性能.在单分类器中决策树 C4.5 的分类效果较好,本文将同时与 C4.5 分类器进行对比,如图 6 所示为 4 种方法(ECDD,C4.5,DDM,STEPT)的分类准确率对比.

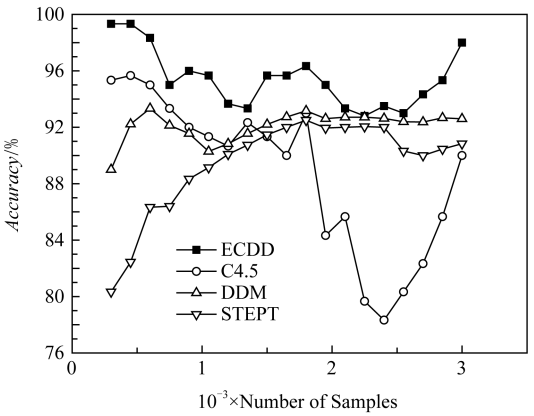


Fig. 6　Comparison of classification accuracy

图 6　4 种方法分类准确率对比

如图 6 所示,在检测到概念漂移时,ECDD 方法可以快速地更新分类器,根据检测到的概念漂移点引入新流量重新学习,更新分类器使分类精度回升.C4.5 方法由于没有概念漂移检测机制,在发生概念漂移时无法引入新流量更新分类器.DDM 与 STEPT 基于分类准确率判断概念漂移的产生,由于流量不同类别的识别率存在差异,因此对于总体的识别准确率有较大差异,导致这 2 种方法可能无法有效地检测概念漂移,最终影响分类准确率.

如图 7 所示为 4 种分类方法(ECDD,C4.5,DDM,STEPT)的综合评价对比.

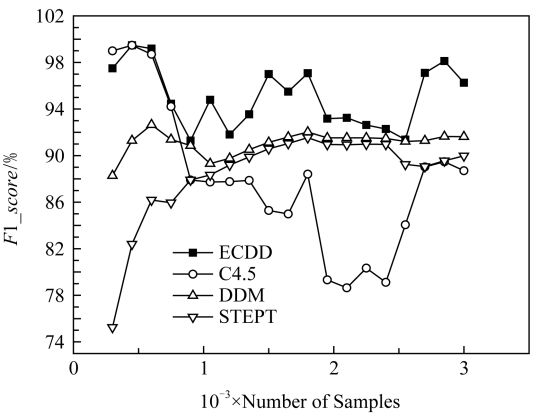


Fig. 7　Comparison of $F1\_score$

图 7　4 种方法综合评价对比

由图 7 可见,ECDD 方法在 $F1\_score$ 上总体优于 C4.5,DDM,STEPT 等方法.

综上,ECDD 方法的分类效果明显好于 C4.5,DDM,EDDM 等方法,ECDD 方法可以及时有效地更新分类模型,使得分类器保持良好的分类性能.

4　结　　论

网络应用层出不穷,不同网络应用可能具有不同的特征分布,导致包含这些应用的网络业务流特征和分布随之发生改变,产生概念漂移问题,概念漂移导致分类模型分类精度降低,性能下降.本文提出一种基于散度的网络流概念漂移分类方法 ECDD,该方法首先根据网络流量特征的 JS 散度变化检测概念漂移,在检测到概念漂移后根据增量集成学习策略引入新样本建立的分类器;接着根据权值比较替换性能下降的分类器;最后,将更新后的集成分类器用于之后的流量分类中.实验结果表明该方法可以有效地检测概念漂移,并在概念漂移点更新分类器,达到良好的分类性能.下一步工作将研究如何处理含噪声的概念漂移网络流.

参　考　文　献

[1] Boukhtouta A, Mokhov S A, Lakhdari N E, et al. Network malware classification comparison using DPI and flow packet headers [J]. Journal of Computer Virology and Hacking Techniques, 2016, 12(2): 69-100

[2] Shafiq M, Yu Xiangzhan, Bashir A K, et al. A machine learning approach for feature selection traffic classification using security analysis [J]. The Journal of Supercomputing, 2018, 74(10): 4867-4892

[3] Bakker J, Ng B, Seah W K G, et al. Traffic classification with machine learning in a live network [C] //Proc of 2019 IFIP/IEEE Symp on Integrated Network and Service Management. Piscataway, NJ: IEEE, 2019: 488-493

[4] Pacheco F, Exposito E, Gineste M, et al. Towards the deployment of machine learning solutions in network traffic classification: A systematic survey [J]. IEEE Communications Surveys & Tutorials, 2018, 21(2): 1988-2014

[5] Žliobaité I. Concept drift over geological times: Predictive modeling baselines for analyzing the mammalian fossil record [J]. Data Mining and Knowledge Discovery, 2019, 33(3): 773-803

[6] Wen Yimin, Tang Shiqi, Feng Chao, et al. Online transfer learning for mining recurring concept in data stream classification [J]. Journal of Computer Research and Development, 2016, 53(8): 1781-1791 (in Chinese)

(文益民, 唐诗淇, 冯超, 等. 基于在线迁移学习的重现概念漂移数据流分类[J]. 计算机研究与发展, 2016, 53(8): 1781-1791)

[7] Pan Wubin. Research on finer-grained classification for encrypted traffic [D]. Nanjing: Southeast University, 2018 (in Chinese)
(潘吴斌. 加密流量精细化分类技术研究[D]. 南京: 东南大学, 2018)

[8] Yu Sun, Ke Tang, Zhu Zexuan, et al. Concept drift adaptation by exploiting historical knowledge [J]. IEEE Transactions on Neural Networks and Learning Systems, 2018, 29(10): 4822-4832

[9] Zhang Dandan, Shen Hong, Tian Hui, et al. A selectively re-train approach based on clustering to classify concept-drifting data streams with skewed distribution [C] //Proc of Pacific-Asia Conf on Knowledge Discovery and Data Mining. Berlin: Springer, 2004: 413-424

[10] Gama J, Medas P, Castillo G, et al. Learning with drift detection [C] //Proc of Brazilian Symp on Artificial Intelligence. Berlin: Springer, 2004: 286-295

[11] Nishida K, Yamauchi K. Detecting concept drift using statistical testing [C] //Proc of Int Conf on Discovery Science. Berlin: Springer, 2007: 264-269

[12] Bifet A, Read J, Pfahringer B, et al. CD-MOA: Change detection framework for massive online analysis [C] //Proc of Int Symp on Intelligent Data Analysis. Berlin: Springer, 2013: 92-103

[13] Nishida K, Shimada S, Ishikawa S, et al. Detecting sudden concept drift with knowledge of human behavior [C] //Proc of IEEE Int Conf on Systems, Man and Cybernetics. Piscataway, NJ: IEEE, 2008: 3261-3267

[14] Borchani H, Larrañaga P, Bielza C. Classifying evolving data streams with partially labeled data [J]. Intelligent Data Analysis, 2011, 15(5): 655-670

[15] Alippi C, Roveri M. Just-in-time adaptive classifiers—Part I: Detecting nonstationary changes [J]. IEEE Transactions on Neural Networks, 2008, 19(7): 1145-1153

[16] Dries A, Rückert U. Adaptive concept drift detection [J]. Statistical Analysis and Data Mining, 2009, 2(5/6): 311-327

[17] Pan Wubin, Cheng Guang, Guo Xiaojun, et al. An adaptive classification approach based on information entropy for network traffic in presence of concept drift [J]. Chinese Journal of Computers, 2017, 40(7): 1556-1571 (in Chinese)
(潘吴斌, 程光, 郭晓军, 等. 基于信息熵的自适应网络流概念漂移分类方法[J]. 计算机学报, 2017, 40(7): 1556-1571)

[18] Baumgartner D, Serpen G. Performance of global-local hybrid ensemble versus boosting and bagging ensembles [J]. International Journal of Machine Learning and Cybernetics, 2013, 4(4): 301-317

[19] Didona D, Romano P. Using analytical models to bootstrap machine learning performance predictors [C] //Proc of the 21st Int Conf on Parallel and Distributed Systems. Los Alamitos, CA: IEEE Computer Society, 2015: 405-413

[20] Hidayati R, Kanamori K, Ling Feng, et al. Combining feature selection with decision tree criteria and neural network for corporate value classification [C] //Proc of the 14th Pacific Rim Knowledge Acquisition Workshop. Berlin: Springer, 2016: 31-42

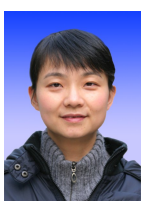
[21] Zhao Yuyu, Cheng Guang, Li Haodong, et al. Active queue management algorithm for time delay demand [J]. Scientia Sinica: Informationis, 2019, 49(10): 1321-1332 (in Chinese)
(赵玉宇, 程光, 李昊冬, 等. 面向时延需求的主动队列管理方法[J]. 中国科学: 信息科学, 2019, 49(10): 1321-1332)

Cheng Guang, born in 1973. PhD, professor, PhD supervisor. His main research interests include network security, network measurement and behavior, future Internet security.

Qian Dexin, born in 1996. Master. His main research interests include network measurement and traffic classification.

Guo Jianwei, born in 1979. Master. His main research interests include service quality modeling and traffic classification.

Shi Haibin, born in 1996. Master candidate. His main research interests include network measurement and network security.



Wu Hua, born in 1973. PhD, associate professor, master supervisor. Her main research interests include network measurement, network management and network security.

Zhao Yuyu, born in 1994. PhD candidate. His main research interests include network measurement and network management.