基于信息熵的自适应网络流概念漂移分类方法

潘吴斌 程 光 郭晓军 黄顺翔

(东南大学计算机科学与工程学院 南京 210096)

(计算机网络和信息集成教育部重点实验室(东南大学) 南京 210096)

摘 要 由于网络流量特征随时间和网络环境的变化而发生改变,导致基于机器学习的流量分类方法精度明显降低.同时,根据经验定期更新分类器是耗时的,且难以保证新分类器泛化性能.因而,文中提出一种基于信息熵的自适应网络流概念漂移分类方法,首先根据特征属性的信息熵变化检测概念漂移,再采用增量集成学习策略在概念漂移点引入当前流量建立的分类器,并剔除性能下降的分类器,达到更新分类器的目的,最后加权集成分类结果.实验结果表明该方法可以有效地检测概念漂移并更新分类器,表现出较好的分类性能和泛化能力.

关键词 概念漂移;机器学习;信息熵检测;增量集成学习;流量分类

中图法分类号 TP393 **DOI 号** 10.11897/SP.J.1016.2017.01556

An Adaptive Classification Approach Based on Information Entropy for Network Traffic in Presence of Concept Drift

PAN Wu-Bin CHENG Guang GUO Xiao-Jun HUANG Shun-Xiang

(*School of Computer Science and Engineering, Southeast University, Nanjing* 210096)

(*Key Laboratory of Computer Network and Information Integration, Ministry of Education, Nanjing* 210096)

Abstract In recent years, traffic classification based on machine learning shows a high accuracy. Nevertheless, machine learning-based traffic classification heavily depends on the environment where the samples are trained. In practice, although a classifier can be accurately trained at a given network environment, its accuracy will see a great decline when it faces to classify traffic from varying network condition in practice. Due to dynamic changes of traffic statistics and distribution, the machine learning-based classifiers should be updated periodically in order to optimize the performance. This issue is unavoidable for machine learning-based traffic classification. The present solutions lack explicit recommendations on when a classifier should be updated and how to effectively update the classifier. These result in several shortcomings: (1) Updating a traditional traffic classifier is time consuming. It is inherent to how often a classifier should be updated or when a new classifier will be needed. (2) Updating only a new classifier on new traffic leads to some learned knowledge lost. It further affects the performance when updating a classifier on a large dataset that combines all collected data. (3) Traffic statistics and distribution from varying network condition are dynamically changed. Thus, it is hard to obtain stable feature subset to build robust classifier. Therefore, building an adaptive classifier to changing network condition is

收稿日期:2015-09-21;在线出版日期:2016-05-10.本课题得到国家"八六三"高技术研究发展计划项目基金(2015AA015603)、江苏省未来网络创新研究院未来网络前瞻性研究项目(BY2013095-5-03)、江苏省"六大人才高峰"高层次人才项目(2011-DZ024)、中央高校基本科研业务费专项资金和江苏省普通高校研究生科研创新计划资助项目(KYLX15_0118)资助.**潘吴斌**,男,1987 年生,博士研究生,主要研究方向为网络安全、网络测量及流量分类.E-mail:wbpan@njnet.edu.cn.**程 光**,男,1973 年生,博士,教授,博士生导师,主要研究领域为网络安全、网络测量与行为学及未来网络安全.**郭晓军**,男,1983 年生,博士研究生,主要研究方向为网络安全、网络测量及网络管理.**黄顺翔**,男,1991 年生,硕士研究生,主要研究方向为网络安全、网络测量及流量分类.

a huge challenge. In this paper, we develop an adaptive traffic classification using entropy-based detection and incremental ensemble learning, assisted with embedded feature selection. In order to update the classifier timely and effectively, the entropy-based detection utilizes sliding window technique to measure the statistical difference between the previous and current traffic samples by counting and comparing all instances with respect to their feature stream membership. Additionally, we discretize the range of feature values to a fixed number of bins to take the approximate value distribution into account. Moreover, incremental ensemble learning schema retains previous trained classifiers, and introduces the classifier retrained on current traffic and removes the classifier with performance degradation. Furthermore, several feature selectors are integrated to obtain feature subsets with robust generalization. The comprehensive performance evaluation conducted on two real-world network traffic data sets shows that our approach can effectively detect concept drift in changing network condition and update the classifier with high accuracy and generalization ability. The major contributions of this work are summarized as follows: first, this paper presents an adaptive traffic classification system based on concept drift detection. Information entropy is used to detect concept drift based on the entropy change of feature attributes. The information entropy-based detection method does not require class information of flows. Second, the classifiers are updated according to the result of concept drift detection, rather than regularly updated at a given period. Third, the method uses ensemble learning strategy to introduce classifier built on new samples, and eliminates classifiers with performance degradation in order to optimize the classification model. Fourth, mutual information is introduced to evaluate features for concept drift detection. The results show that the mutual information between packet size and protocol is high and stable, which indicates that the feature is suitable for concept drift detection. Fifth, this paper uses Hoeffding boundary to determine the window size of concept drift detection. The appropriate window size is significant for fast and effective concept drift detection.

Keywords　concept drift; machine learning; information entropy detection; incremental ensemble learning; traffic classification

1　引　言

随着移动互联网的快速发展，网页浏览、流媒体以及社交网络中新业务不断出现，同时用户网络安全需求使得加密流量占比不断增加，使得传统的流量分类方法面临严峻的挑战. 针对 DPI(Deep Packet Inspection)分类方法[1]解析数据包负载内容侵犯隐私，且对加密业务无能为力，促使研究人员转向基于机器学习的流量分类方法[2-5]. 但基于流特征的机器学习分类方法会因为不同时间段以及不同地域的流量所承载的业务分布差异而引起概念漂移问题[6-7]. 因此，根据先前流量训练的分类器对新样本空间的适用性逐渐变弱，导致分类模型的识别能力下降. 针对该问题研究人员进行了深入研究，但还存在一些不足：(1) 只在新的流量上重新训练分类器，导致一些历史知识丢失，而且重新标记样本代价高；(2) 结合不同时期收集的所有流量训练分类器会导致性能问题[8]. 此外，如果某个特定时期具有较大的数据量，将对流量分类起主导作用. 为了避免这种情况，需要从不同时期选择代表性样本用来构成复合数据集；(3) 定期频繁更新分类器不仅耗费时间和资源，且难以保证分类器泛化能力，如何显式发现概念漂移有利于更新分类器；(4) 随着网页浏览和流媒体中新业务的不断出现，无法收集和分析完整的训练样本，训练样本的数量及质量对识别性能具有较大的影响. 因此，构建自适应复杂多变网络环境的分类模型是一个巨大的挑战.

针对上述问题，本文借鉴信息论和集成学习思想，提出一种基于信息熵的自适应网络流概念漂移分类方法. 首先，该方法根据特征属性分布的熵变化检测概念漂移，然后，借助增量集成学习策略在保留

原来分类器的基础上,在概念漂移点引入新流量训练的分类器,根据精度权重替换原有性能下降的分类器,使得分类器得到有效更新.本文采用两组网络流数据集验证该方法在网络流突变和渐变时的分类性能,实验结果表明该方法在应对概念漂移问题时具有较高的准确率和泛化能力.本文研究贡献主要包括以下几点:

(1)本文提出基于概念漂移检测的自适应流量分类系统.第一,采用信息熵根据特征属性分布的熵变化检测概念漂移,与性能检测方法不同的是基于信息熵的概念漂移检测方法不需要流量的类别信息.第二,根据概念漂移检测结果更新分类器,不再根据固定的周期定时更新分类器,保证分类器的有效更新.该方法采用集成学习策略在保留原分类器的基础上引入新样本构建的分类器,并剔除分类性能下降的分类器,而不是集成所有新样本,有利于优化模型,提高分类性能和效率.第三,该方法只需首次标记训练样本,后引入的新样本由分类器自动标记,无需再次人工标记.

(2)本文引入互信息评估概念漂移检测特征.当前网络流量特征多,选择稳定的、计算量小的检测特征对改善概念漂移检测性能至关重要.由于包级特征计算量最小,有利于概念漂移检测快速进行.并采用互信息评估各包级特征包含协议类型的信息量,结果显示包大小特征包含信息量高且稳定,表明该特征适用于概念漂移检测.

(3)本文根据 Hoeffding 边界确定概念漂移检测窗口大小的阈值.选择合适的窗口大小对于快速有效地检测概念漂移意义重大.窗口过大增加检测延迟,且容易造成漏报;窗口过小受噪声影响大,容易造成误报.

本文第 2 节综述网络流概念漂移的相关研究;第 3 节详细论述流量分类中的概念漂移问题;第 4 节阐述基于信息熵的自适应流量分类系统;第 5 节给出实验数据集、简要说明实验环境,并对分类算法性能进行分析;第 6 节总结全文并展望未来的工作.

2 相关研究

自"概念漂移"在1986年由Schlimmer和Granger[9]首次提出后,研究人员对概念漂移问题进行了大量研究,并取得了一定的成果[10].关于概念漂移的解决方法主要有两种:(1)增量更新.该方法不能明确地检测概念漂移,它随新样本的到来增量地更新分类器,该方法的目标是提高准确性.Zhong 等人[11]在 CVFDT[12]的基础上提出了解决 P2P 流量概念漂移问题的算法 iCVFDT,该算法在发生概念漂移时生成一棵新子树,新子树构建完成后替换原来的子树,从而解决概念漂移带来的性能下降,但该方法并未考虑多种应用情况下的概念漂移.Erman 等人[13]将半监督学习方法引入流量分类,但该方法采用的聚类算法本身分类准确率不高,并缺乏与其他算法性能的比较.根据半监督学习思想,Raahemi 等人[14]提出基于 tri-training 的 P2P 流量分类,采用滑动窗口技术周期性评估窗口样本的分类准确率,以此检测概念漂移,该方法周期性性能评估和重新训练模型需要耗费大量时间和资源,无法用于大规模网络及多类识别;(2)检测和再更新.该方法明确地检测概念漂移并在概念漂移被发现时更新分类器.该方法优点是针对概念漂移显性更新分类器.Gama 等人[15]提出根据分类错误率检测概念漂移的方法 DDM,该方法持续地监视分类错误率.如果分类错误率低于阈值就报告发生概念漂移.该方法对于突变式的概念漂移表现优越,但不善于检测渐变式的概念漂移.Nishida 等人[16]提出了基于统计的检测方法 STEPD,该方法与 DDM 具有相似的架构,但该方法采用统计检验来检测概念漂移,通过比较全局准确性和近期准确性来识别概念漂移,该方法缺点是需要一个预先给出滑动窗口的大小.如果大小不合适,容易产生误报或漏报.上述方法采用样本的局部信息,延缓了概念漂移的识别.例如,DDM 只用了不正确分类的信息,而 STEPD 只用了正确分类的信息.针对这个问题,Bifet 等人[17]用一个可变的滑动窗口来检测随时间变化的数据流的方法 ADWIN,滑动窗口将分类结果作为输入,且窗口大小持续增长.无论新样本何时进入,ADWIN 都将窗口分裂成两个子窗口,并比较 $|u_1-u_2|$(u_1 和 u_2 是两个子窗口的平均值)和 Hoeffding 界,当 $|u_1-u_2|$ 超过 Hoeffding 界时表示发生概念漂移.当概念漂移被发现时,窗口收缩,该方法缺点是窗口增长所需的时间长.Du 等人[18]使用滑动窗口的信息熵来检测概念漂移,滑动窗口大小动态地从 Hoeffding 界获取,该方法只利用正确和不正确分类的类别属性,不适用于多类问题.Lu 等人[19]提出一种基于案例推理系统检测概念漂移的方法,引入检测胜任力变化的新胜任力模型,而不是统计实际分布情况.该方法不需要先验分布的知

识，且概念漂移检测可靠性高.

针对网络流概念漂移给流量分类带来的不利影响，本文提出一种基于信息熵的自适应分类方法. 该方法根据特征属性分布的信息熵变化检测概念漂移，并根据增量集成学习策略引入新流量更新分类器，增量集成学习可以保留先前训练的分类器，并剔除分类性能下降的分类器，最后根据加权分类器集成分类结果，有效应对网络流概念漂移问题.

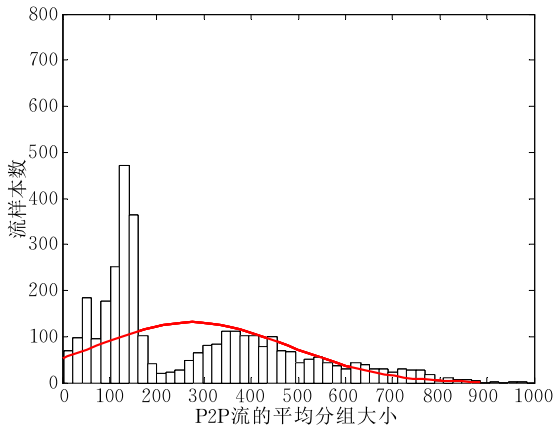
3　网络流概念漂移

网络中大规模流量数据包高速传输，网络流量是一种典型的数据流应用. 由于网络流分布随网络环境动态变化，如不同时间段网络行为的差异，以及不同地域流量承载的业务的差异，使得业务分布发生较大的变化，导致基于流特征的机器学习分类方法识别能力会因此而下降，分类模型需要频繁的重建，这是数据流挖掘中典型的概念漂移问题[4,11]. 从概率角度来看，分类器是通过计算流特征 $X=\{x_1,x_2,\cdots,x_n\}$ 被分为 Y 的概率来建立的. 由于 X 是已知的，分类结果依赖于概率 $P(y|X)=\dfrac{P(y)|P(X|y)}{P(X)}$. 分类器可以定义为期望函数 $f:X\to Y$：

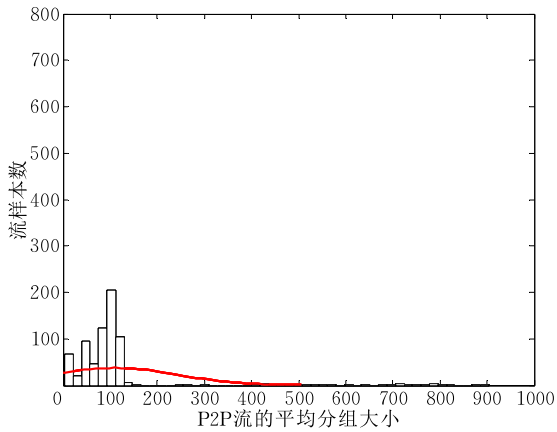
$$f(X)=\arg\max_{y\in Y}P(y|X)=\arg\max_{y\in Y}\frac{P(y)\prod_{i=0}^{n}P(x_i|y)}{P(X)}\tag{1}$$

从式(1)可以看出，分母 $P(X)$ 是样本统计特征 X 的概率，$P(X)=\prod_{i=0}^{n}P(x_i)$，$P(x_i)$ 是 x_i 在给定训练集的概率，对于所有类别都是常数；然而，流统计特征 x_i 变化会导致独立概率分布 $P(x_i|y)$ 变化，从而影响 $P(y|X)$；同时，类别先验概率 $P(y)$ 变化也会影响 $P(y|X)$.

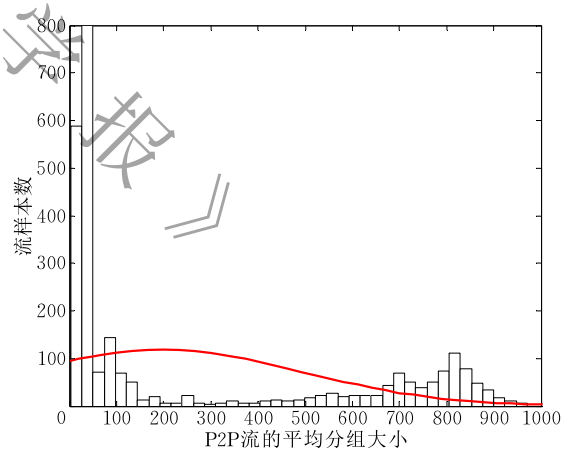
为了描述流量特征的变化，以 UNIBS 数据集[20]中 P2P 流量的平均分组大小为例描述连续 3 天的 P2P 流分布变化，如图 1 所示. P2P 流包含 Edonkey 和 Bittorrent 两种不同的 P2P 应用，图 1 中 Day1，Day2，Day3 分别包含 3500、700 和 3800 个 P2P 样本，Day1 与 Day2 相比较，Day2 平均分组大小主要集中在 0～150 字节，超过 150 字节的占比很小；Day1 与 Day3 相比较，Day3 平均分组大小主要集中在 0～100 字节，超过 100 字节的占比较小.



(a) 第1天的P2P流分布



(b) 第2天的P2P流分布



(c) 第3天的P2P流分布

图 1　UNIBS 数据集连续 3 天的 P2P 流分布

4　基于信息熵的自适应分类方法

如果能够准确地识别概念漂移，就可以及时有效地更新分类器，从而避免仅根据经验设置固定的时间间隔频繁更新分类器. 当前主要通过分类准确率的下降来判断是否发生概念漂移，但基于分类准确性的概念漂移检测需要类别标记，与预测类别相

比较以统计分类准确率,然而,标记样本需要耗费大量时间和资源.另外,基于分类准确性的概念漂移检测很难应用于类不平衡的网络流,对于占大多数的协议概念漂移检测不存在问题,但对于占少数的协议,误差率变化大,检测性能不稳定.

为了解决基于分类准确性的概念漂移检测方法的不足,本文提出一种基于熵检测的自适应分类方法(Adaptive Classification based on Entropy Detection,ACED).该系统主要包括流统计特征处理、特征选择模块、概念漂移检测机制和集成分类器,如图 2 所示.流统计特征处理用于组流,并获取流的统计特征;特征选择模块用于选取建立分类器的稳定的特征子集;概念漂移检测机制用于检测流量是否发生概念偏移,确定何时更新分类器;集成分类器通过综合多个加权分类器的分类结果识别新流量,并在概念漂移点引入新流量更新分类器,构建出适应新环境的分类模型.

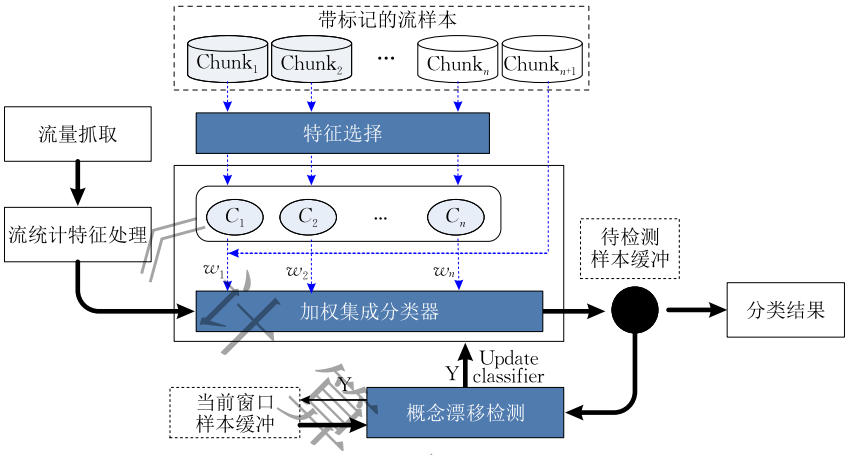


图 2 ACED 系统架构

4.1 基于熵的概念漂移检测机制

针对随时间推移以及不同网络环境引起的流特征变化导致的概念漂移问题,分类器对新流量的适应能力将逐渐变差,很难保持稳定的识别率和泛化能力.如果能够准确地识别概念漂移,就可以及时有效地更新分类器,不然只能根据经验设置固定的更新间隔,频繁地更新分类器需要消耗大量的时间和资源.当前基于分类准确性的概念漂移检测方法在网络流量检测中由于类不平衡性问题不能很好地适用.另外,在流量识别过程中,评估分类错误率所需的类型属性信息需要人工标记,很难获取.针对以上问题,本文根据流特征属性分布变化提出一种基于信息熵的网络流概念漂移检测方法,熵是一个集合的信息量的度量,为了在网络流环境下使用该度量,采用滑动窗口技术比较两个窗口,一个代表数据流中旧一些的实例,另一个代表新的实例.该方法并不直接比较两个窗口的熵,而是将流特征属性离散化为若干个分支,并将多个流特征一同比较,通过统计各特征属性以及各分支的熵来比较两个窗口的差异.

定义流在 t_i 时刻为 $\boldsymbol{d}_i$,$\boldsymbol{d}_i$ 包含 S 个特征的特征集 $\boldsymbol{s}$ 和标记 l,$\boldsymbol{d}_i=(\boldsymbol{s}_i,l_i)$,由于香农熵 $H(x)=-\sum_x P(x)\log_2[P(x)]$,对应 t_i 时刻的熵

$$H_i=\frac{1}{S}\sum_{s=1}^{S}H_{is} \tag{2}$$

S 表示特征的数目,因此,

$$H_{is}=\sum_{b=1}^{B}H_{isb} \tag{3}$$

H_i 代表滑动窗口 $I(s_i,Y)$ 时刻在分支($b\in B$)和特征($s\in S$)上的熵,B 代表各特征的分支,H_{isb} 代表 t_i 时刻在分支 b 和特征 $I(t_i,t_{i-1},Y)$ 上的熵

$$H_{isb}=-w_{isb}[\rho_{isb_{\text{old}}}\log_2(\rho_{isb_{\text{old}}})-\rho_{isb_{\text{new}}}\log_2(\rho_{isb_{\text{new}}})] \tag{4}$$

$\rho_{isb_{\text{old}}}$ 代表 old 窗口 t_i 时刻在分支($b\in B$)和特征($s\in S$)上的概率,

$$\rho_{isb_{\text{old}}}=\frac{v_{isb_{\text{old}}}}{\lambda_{i_{\text{old}}}},\quad \rho_{isb_{\text{new}}}=\frac{v_{isb_{\text{new}}}}{\lambda_{i_{\text{new}}}} \tag{5}$$

w_{isb} 代表每个分支的权重,$\sum_{s=1}^{S}\sum_{b=1}^{B}w_{isb}=1$,当窗口大小 $\lambda_{i_{\text{old}}}=\lambda_{i_{\text{new}}}$,为了简化计算,令 $w_{isb}=1$.

4.2 检测窗口大小

基于信息熵的检测方法通过比较滑动窗口的熵值来实现,窗口大小过大,容易造成漏报,且带来较大的检测延迟,窗口大小过小,噪声影响会较大,带来较多的误报,因此,确定合适的窗口大小是首要任

务.本文窗口大小阈值 ξ 根据 Hoeffding 边界[21]确定.

Hoeffding 边界描述如下：随机变量 R 的 n 个独立样本的均值和真实平均值的误差不超过 ε 的概率为 $1-\delta$，可得

$$\varepsilon=\sqrt{\frac{R^2\ln(1/\delta)}{2n}} \tag{6}$$

鉴于 Hoeffding 边界 ε 随 n 的增加而减小，当 n 增长到足够大时，Hoeffding 边界 ε 足够小，即当前节点进行分裂的最小取值.根据 Hoeffding 边界可以得到最小窗口大小 ξ：

$$\xi=\frac{R^2\ln(1/\delta)}{2\varepsilon^2} \tag{7}$$

假设两个独立样本集来自于相同的随机变量，u_1，u_2 分别代表两个样本集的均值，真实平均值为 u_0.根据 Hoeffding 边界可得 $|u_0-u_2|\leq\varepsilon$，$|u_0-u_1|\leq\varepsilon$，可得

$$|u_1-u_2|\leq 2\cdot\varepsilon \tag{8}$$

因此，由式(7)和(8)可得

$$\xi=\frac{2R^2\ln(1/\delta)}{(u_1-u_2)^2} \tag{9}$$

Hoeffding 边界默认 $\delta=10^{-7}$，两类问题的随机变量 $R=\log_2(2)=1$.为了解决 u_1-u_2 恒小于 Hoeffding 边界的问题，Domingos 提出 Tie-breaking 方法，实验验证将 ε 设为 0.05 比较合适[22]，即 $|u_1-u_2|\leq 0.1$，因此，可得最小检测窗口 $\xi=1400$.

4.3　概念漂移检测特征

为了选择有效的概念漂移检测特征，采用统计测度来评估不同包级特征所包含的信息量.从每条流的前 n 个分组提取特征，不考虑载荷为空的数据包，空载荷数据包常用于传递连接状态信息，如应答接收到的数据或保持会话连接，因此，对于不同的应用空载荷数据包所起的作用不同，比如数据块和空载荷 ACK 数据包的到达时间间隔传达的是 TCP 状态信息，而不是应用层协议的控制信息.TCP 流只考虑前 3 个分组中有 ACK 应答的双向流，$S=(s_1,s_2,\cdots,s_n)$，s_i 代表第 i 个分组的负载大小；$T=(t_1,t_2,\cdots,t_n)$，t_i 是第 i 个分组和第 $i+1$ 个分组之间的到达时间间隔，$t_i=10\lg(I_i/1\,\mu s)$，I_i 表示第 i 个分组实际到达时间间隔；$D=(d_1,d_2,\cdots,d_n)$，d_i 是第 i 分组的传输方向，如果与上一个分组方向相同，则 $d_i=1$；否则 $d_i=-1$.采用互信息 $I(X,Y)$ 评估流特征 X 与协议类别 Y 的关系，估计每个特征包含协议类别的信息量.$I(X,Y)$ 用于评估特征 X 与协议类别 Y 的依存关系，如果特征 X 与协议类别 Y 不相关，$I(X,Y)$ 趋向于 0，$I(X,Y)$ 越大表明相关性越强.因此，对流特征(负载大小、到达时间间隔和分组传输方向)分别进行互信息统计.具体统计流前几个包的包级特征所包含协议类别的信息量，用互信息[23]来表示：

$$I(X,Y)=\sum_{x\in X}\sum_{y\in Y}p(x,y)\log\frac{p(x,y)}{p(x)p(y)} \tag{10}$$

$p(x,y)$ 表示联合分布，$p(x)$，$p(y)$ 表示边缘分布，互信息是非负的，越大表明相关性越强.

4.4　概念漂移检测算法

该方法将问题简化为两个样本集的比较，如果熵 $H_i>\tau$，称 t_i 时刻为概念漂移点.算法 1 描述了概念漂移检测方法.算法实际上为 k 次独立的计算，每次为一个三元组 $(\lambda_{i_{\text{old}}},\lambda_{i_{\text{new}}},H_i)$，函数 H_i 为两组样本的差异性.$Win_{1,i}$ 是基窗口，包含从上次检测到概念漂移点后开始的 $\lambda_{i_{\text{old}}}$ 个样本，$Win_{2,i}$ 是最新的 $\lambda_{i_{\text{new}}}$ 个样本，当新样本加入时，窗口 $Win_{2,i}$ 向前滑动.每次更新，检测是否 $H_i>\tau$，如果是，报告发生概念漂移，重复整个过程.

算法 1.　概念漂移检测算法.

```
1. Initialize: c0=0, SlideSize=constant
2.    Win1,i=first λ_iold points from time c0
3.    Win2,i=next λ_inew points in flows
4. While not at end of flow do
5.    for i=1…k do
6.      Slide Win2,i by SlideSize points
7.      if H(Win1,i,Win2,i)>τ then
8.        c0=current time
9.        Report change at time c0 and update classifier
10.       Clear all windows and GOTO step 2
11.     end if
12.   end for
13. end while
```

4.5　分类器特征选择及基分类器

网络流特征会随时间和环境变化产生概念漂移现象，而分类模型的特征子集都是基于给定数据集获取的，该特征子集无法建立有效的分类模型适应未来很长一段时间的分类要求.此外，流特征中的冗余和不相关特征需要剔除，从而简化分类模型，提高分类模型的泛化能力及分类性能.比如端口号、流持续时间和平均报文大小的区分性明显，而平均报文数、实际接收到的字节数等特征针对某些应用可能区分性不足.因此，急需有效的特征选择方法剔除冗余和不相关特征获取高泛化能力的特征子集，维持较好的分类效果和效率.

本文针对流量概念漂移问题,采用基于选择性集成策略的混合式特征选择方法 FSEN[24] 选取稳定的特征子集.混合式特征选择方法 FSEN 主要由两部分组成,第一部分,根据选择性集成策略从一组特征选择器中选择部分集成,理论分析和实验验证[24]表明集成部分特征选择器优于集成所有特征选择器.为了提高混合特征选择方法的泛化能力,选取相关性、距离、信息增益和一致性 4 种度量的特征选择算法集成,分别为 FCBF、Chi-square、InfoGain、GainRatio 和 Consistency.然后,将这些特征子集根据评价指标进行排序,剔除不满足特定指标的特征,将满足要求的特征子集进行集成.第二部分采用 Wrapper 方法每次从未选择的特征中选择一个特征 x 加入特征子集 X,并且将该特征加入与已选特征进行组合,采用朴素贝叶斯算法评估此时的特征子集,当增加特征时分类准确率下降,则该过程终止,以此获取全局最优特征子集,流程如图 3 所示,详细分析可参见文献[24].

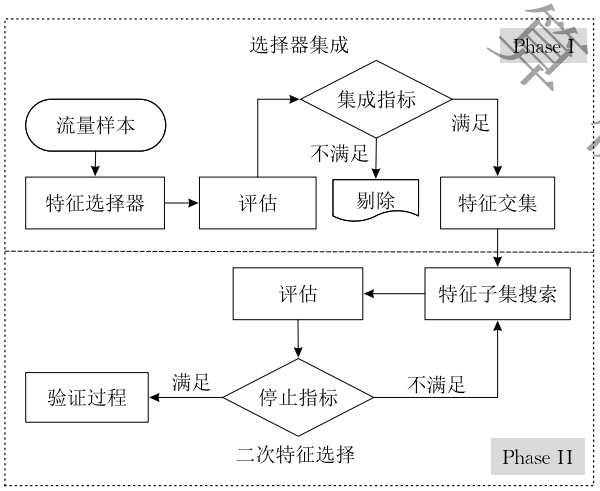


图 3 混合特征选择 FSEN 流程图

选取 C4.5 决策树作为集成学习的基分类器,因为 C4.5 决策树相较于其他机器学习算法具有较好的分类性能:(1) C4.5 的分类模型易于理解,且具有较好的分类效果和效率;(2) C4.5 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.

4.6 增量集成学习

为了有效更新分类器,我们采用增量集成学习更新方式,一方面,充分利用先前训练的分类器,另一方面,在保留先前训练的分类器的基础上引入当前样本训练的分类器集成,并剔除性能下降的分类器,保证集成分类器的泛化能力.该方法将训练集分成大小相同的块 $S_1,S_2,\cdots,S_n$,其中 S_n 表示最近的数据块.C_i 表示从训练数据集 S_i 学习得到的分类器,G_k 表示从整个数据集最后 k 个数据块 $S_{n-k+1}\cup\cdots\cup S_n$ 学习得到的分类器,E_k 表示由最后 k 个分类器 $C_{n-k+1},\cdots,C_n$ 集成获取的分类器.在概念漂移环境中,将之前学习得到的模型应用于分类当前的测试集可能存在显著的偏差.因而,本文基于均值的集成学习方法,提出了一种新的基于权重的方法.该方法给每一个分类器 C_i 一个权重 w_i,其中 w_i 反比于 C_i 分类当前测试集的期望误差.Wang[25] 证明如果给每一个分类器分配权重,给 E_k 中的每一个分类器根据分类器在测试集上的期望分类准确率分配权重,则 E_k 比 G_k 产生更小的分类错误率.这意味着与单个分类器 G_k 相比,如果给集成分类器中的每一个分类器分配一个权重并且权重反比于其期望误差,集成分类器可以减少分类误差.

将训练集分成大小相同的块 $S_1,S_2,\cdots,S_n$,其中 S_n 是最新的块.从每一个 S_i 中学习一个分类器 $C_i,i\geq 1$.对于给定的测试集 T,赋予每个分类器 C_i 一个权重,这个权重与 C_i 分类测试集 T 的期望误差成反比.获取分类器 C_i 的权重通过评估其在测试集的期望预测误差.假设最近的训练数据集 S_n 的类别分布是最接近当前测试集的类别分布,因而分类器的权重可以通过计算分类器在 S_n 的分类误差来近似估计.具体地说,假设 S_{n+1} 是由 (x,c) 数据格式构成,其中 c 是当前记录的真实标记,C_i 对实例 (x,c) 的分类错误率为 $1-f_c^i(x)$,$f_c^i(x)$ 是分类器 C_i 判断样本 x 标记为类别 c 的概率.因而,分类器 C_i 的均方差为

$$MSE_i=\frac{1}{|S_n|}\sum_{(x,c)\in S_n}(1-f_c^i(x))^2 \tag{11}$$

分类器 C_i 的权重反比于 MSE_i.另外,随机分类器的分类错误率 $p(c)$ 的均方差为

$$MSE_r=\sum_c p(c)(1-p(c))^2 \tag{12}$$

由于随机分类器并不包含样本的有用信息,根据随机分类器的错误率 MSE_r 判断是否加入集成分类器.如果分类器的错误率小于 MSE_r,则加入集成分类器;否则丢弃.集成分类器中各分类器 C_i 的权重 w_i 计算如下:

$$w_i=MSE_r-MSE_i \tag{13}$$

算法 2 描述了基于权重的集成学习算法的流程:第 1~2 行从最近的数据集 S 中获取分类器 C',并计算分类器 C' 的权重 w';第 3~7 行将 S 作为测试集计算 $C=\{C_1,C_2,\cdots,C_k\}$ 每个初始分类器的权重 $w_i(1\leq i\leq k)$,淘汰 $w_i\leq 0$ 的分类器;最终从 $C\cup$

$\{C'\}$中返回权重前 k 个分类器.

算法 2.　基于精度权重的集成学习方法.

输入：S：从标记文件中获取最近的数据块

k：分类器的数目

C：k 个预先训练的分类器

输出：C：带有更新权重的 k 个分类器的集合

1. 从训练样本集 S 训练分类器 C'；
2. 计算 C' 错误率，获取 C'的权重 w'；
3. for $C_i \in C$ do
4. compute MSE_i；/*将 C_i应用于 S */
5. compute $w_i = MSE_r - MSE_i$；
6. if $w_i \leq 0$ /*淘汰权重 $w_i \leq 0$ 的分类器*/
7. 从 C 中移除 C_i；
8. $C = Topk(C \cup \{C'\})$；/*获取权重前 k 的分类*/
9. return C；

4.7　ACED 算法流程伪代码

算法 3 描述了 ACED的学习和分类过程：行 1～3 是分类过程，输出多个加权分类器的置信度和最大的分类结果；行 4～20 根据信息熵检测概念漂移，并在概念漂移点更新分类器，行 5 检测是否发生概念漂移，具体描述如算法 1；行 6～10 建立分类器 C_{n+1}，并计算分类器 C_{n+1}的权重 w_{n+1}；行 11～14 计算各分类器 $C=\{C_1, C_2, \cdots, C_n\}$的权重 $w_i (1 \leq i \leq n)$；行 15～20 判断分类器 C_{n+1}的权重是否 $w_{n+1} > \min_{i=1}^{n} w_i$，如果是，分类器 C_{n+1}替换权值最小的分类器 C_i，最终返回权重前 M 的分类器集成，并更新分类器.

假设在大小为 s 的数据集上构建一个分类器的复杂度为 $f(s)$，为了获取分类器的权重 w，需要每个分类器分类测试集，而分类测试集的复杂度与测试集的大小成线性关系. 假设整个数据流被分成 n 份，由于熵检测发现概念漂移的次数不确定，单从分类器更新时间复杂度考虑，算法 3 的时间复杂度为 $O(n \times f(s/n) + Ms)$.

算法 3.　ACED 方法.

输入：S：块大小，M：集成分类器的分类器个数

初始化：待检测样本缓冲$(DB)=\varnothing$，

当前窗口样本缓冲$(FB)=\varnothing$

1. While $flow(x_t, y_t)$ is available do
2. get classifers output $\forall_y \forall_i \{H_i^y(x_t)\} \in [0,1]$
3. Output $H(x_t) = \arg\max_{y \in Y} \sum_{i=1}^{n} w_i H_i^y(x_t)$
4. add (x_t, y_t) to DB
5. if $H_t(DB, FB) > \tau$ then
6. add (x_t, y_t) to FB, initialize FB
7. build batch classifier C_{n+1} from FB
8. MSE_{n+1} = error rate of C_{n+1} via the cross validation method on FB；
9. MSE_r = error rate of the random classifiers on FB；
10. $w_{n+1} = MSE_r - MSE_{n+1}$；
11. for $C_i \in C$ do
12. $MSE_i = \frac{1}{|S_n|} \sum_{(x,c) \in S_n} (1 - f_c^i(x))^2$；
13. $w_i = MSE_r - MSE_i$；
14. end for
15. if $n < M$ then
16. $n = n + 1$
17. else if $w_{n+1} > \min_{i=1}^{n} w_i$ then
18. replace C_i with C_{n+1}；
19. end if
20. end if
21. end while

5　实验与分析

5.1　实验数据集

目前，从已有公开的数据集来看，还未出现标准的数据集用于算法性能的评估比较，因此，采用不同数据源的数据集有利于算法性能分析及有效性验证. 本文采用 WIDE[26] 和 Auckland[23] 合成数据集 WAND 模拟不同环境导致的概念漂移，WIDE 和 Auckland 数据集都是载荷被消除的匿名网络流数据，将数据包头部指出的实际报文长度作为分组大小特征. 为了有效地标记该网络流，我们先基于端口号进行标记，然后利用一些过滤规则过滤不完整流量. 针对 HTTP，过滤服务器端发出的报文存在空载荷的流，因为 HTTP 作为 web 应用，服务器发出的报文有负载内容. 针对 SMTP 和 FTP，由于保持连接需要发送一些空载荷的控制报文，因此过滤双向都有空载荷的流. 另外，报文数量太少的流也被过滤，因为报文太少不利于流量分析. WIDE 网络流共包含 368 426 个完整的网络流样本，而 Auckland 网络流共包含 293 203 个完整的网络流样本，被分为 6 种应用类型，如表 1(a)所示. CNT 数据集是采用 tcpdump 抓取华东北网络中心不同网段的双向全报文数据，CNT 数据集共包含 123 280 个完整的双向流网络流样本，报文带有负载内容，采用 ndpi 识别工具进行标记，分为 6 种应用类型，具体分布如表 1(b)所示.

表 1 WAND 和 CNT 流分布

(a) WAND 流分布

数据集	Source	HTTP	SSL	DNS	SMTP	FTP	POP3
WAND	WIDE	275074	62101	24175	6433	399	244
	Auckland	126110	50696	82565	32360	318	1154

(b) CNT 流分布

数据集	Source	HTTP	Flash	SSL	ICMP	QQ	BT
CNT	site A	71500	3935	1818	365	263	147
	site B	40206	1748	2271	213	654	160

本文实验平台为 Windows 7,CPU 为酷睿 i5-3210,内存为 8GB,基于 Java 调用 Weka-3.7.10 开发,开发环境为 eclipse-4.2.2.

5.2 评估指标

准确率是算法性能评估最常用的指标.假设 N 表示样本数量,应用种类用 m 表示.真正 TP 表示应用 i 的样本被正确标记的数目,$TP_i=n_{ii}$.真负 TN 表示应用为非 i 的样本被正确标记的数目,$TN_i=n_{jj}$.假负 FN 表示应用 i 的样本被错误标记的数目,$FN_i=\sum_{j\neq i}n_{ij}$.假设 FP 表示应用为非 i 的样本被错误标记的数目,$FP_i=\sum_{j\neq i}n_{ji}$.根据上述评估参数,给出整体准确率($OA$)、查准率($Precision$)、查全率($Recall$)和综合评价($F$-Measure)的数学描述[24].

$$OA=\frac{\sum_{i=1}^{m}TP_i}{\sum_{i=1}^{m}(TP_i+FN_i)} \tag{14}$$

$$Precision=\frac{TP_i}{TP_i+FP_i} \tag{15}$$

$$Recall=\frac{TP_i}{TP_i+FN_i} \tag{16}$$

$$F\text{-Measure}=\frac{2\times presion\times recall}{presion+recall} \tag{17}$$

整体准确率代表分类方法的总体性能,可以采用查准率和查全率评估算法在各应用上的识别性能,也可以综合查全率和查准率采用 F-Measure 来表示算法在各应用上的识别性能.

为了评估概念漂移检测方法的有效性,采用误报率(FPR)和漏报率(FNR)来评估.误报率和漏报率越高,检测性能越差.另外,采用平均延迟($delay$)来表示检测到概念漂移与实际概念漂移的延迟的平均.

$$FPR=\frac{FP_i}{TN_i+FP_i} \tag{18}$$

$$FNR=\frac{FN_i}{TP_i+FN_i} \tag{19}$$

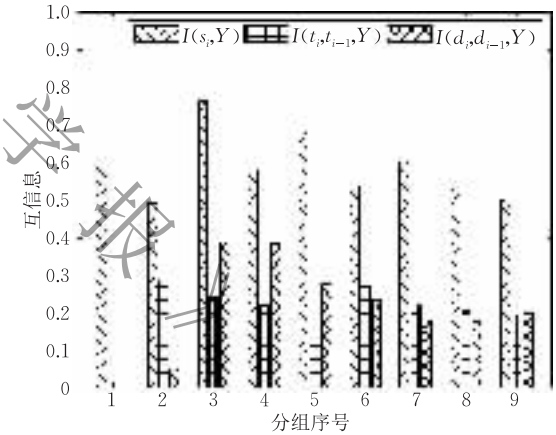
$$delay=\frac{\sum delay \text{ at each drift}}{\text{total number of drift}} \tag{20}$$

5.3 概念漂移检测结果

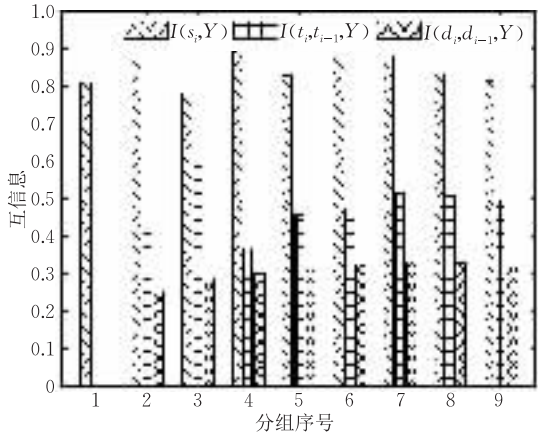
为了有效评估信息熵检测方法,文中选用 2 组不同类型的流量概念漂移数据验证该方法的有效性,其中,WAND 数据集包含 4 个概念,每个概念代表一种网络环境的流量数据,每个概念中包含 10000 个样本(WIDE 数据源在每个概念中占比依次为 80%、20%、70%、30%),属于突变类型;而 CNT 数据集包含 6 个概念,每个概念包含 8000 个样本(Site A 数据源占比依次为 100%、80%、60%、40%、20%、0),属于渐变类型.

5.3.1 检测特征互信息评估

为了进行有效的概念漂移检测,需要选择稳定的、计算复杂度低的检测特征.包级特征提取计算量小,且可以有效地包含流的信息[23].因此,可以采用包级特征进行熵检测.本文采用互信息评估包级特征所包含协议类型的信息量,包级特征包括分组大小、到达时间间隔和传输方向,与协议类别的互信息如图 4 所示.



(a) 数据集CNT



(b) 数据集WAND

图 4 包级特征与协议类别的互信息

图 4 可以看出分组大小与协议类别 Y 的互信息 $I(s_i,Y)$最大，CNT 数据集的互信息 $I(s_i,Y)$为 0.5～0.8，且保持稳定，而到时时间间隔和传输方向与协议类别 Y 的互信息 $I(t_i,t_{i-1},Y)$和 $I(d_i,d_{i-1},Y)$明显小于 $I(s_i,Y)$. WAND 数据集的互信息 $I(s_i,Y)$达到 0.8～0.9，包到达时间间隔特征与协议类别 Y 的互信息 $I(t_i,t_{i-1},Y)$明显低于 $I(s_i,Y)$，互信息 $I(t_i,t_{i-1},Y)$约为 0.2～0.4，而互信息 $I(d_i,d_{i-1},Y)$低于 0.1. 综合来看，包到达时间间隔的互信息相对稳定，但相关性较低，该特征不仅受网络状况影响，还受 QoS 机制影响，不同协议的数据包赋予不同的优先级. 而传输方向特征稳定性较差，在 CNT 数据集中，传输方向特征的互信息与到达时间间隔特征的互信息相差不大，而 WAND 数据集中传输方向特征的互信息明显低于包到达时间间隔特征的互信息. 因此，采用稳定性和相关性较强的分组大小特征作为概念漂移检测特征.

5.3.2　熵检测的阈值

概念漂移检测的阈值影响检测算法的性能，为了合理的选择检测阈值 τ，采用误报率和漏报率来评估不同阈值的检测性能，并采用平均延迟来描述检测到概念漂移与实际概念漂移的延迟，结果如表 2 所示.

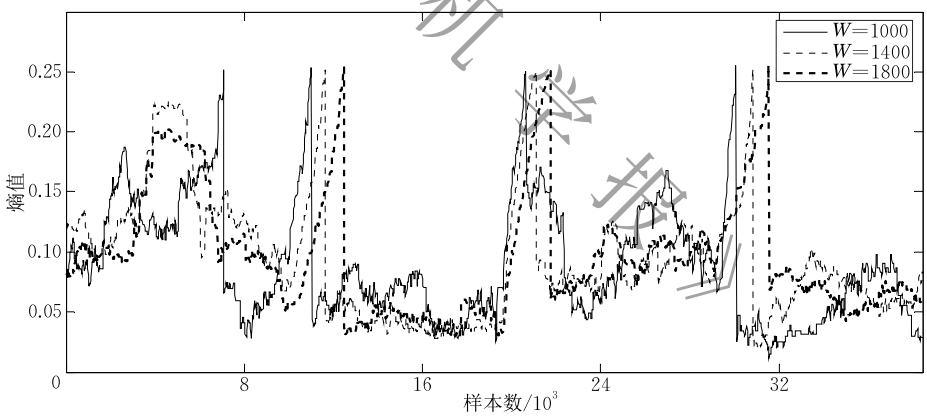
表 2　熵检测的阈值

阈值	WAND			CNT		
	误报	漏报	平均延迟	误报	漏报	平均延迟
0.15	66.7	0	1310	0	0	560
0.20	33.3	0	1545	0	0	1010
0.25	0	0	1700	0	0	1280
0.30	0	33.3	2320	0	60	1410
0.35	0	100.0	—	0	100	—

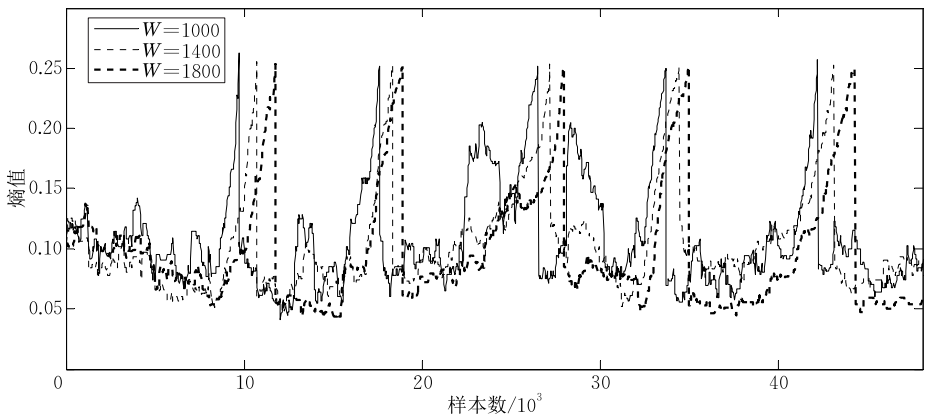
从表 2 可以看出，随阈值增加、误报降低、漏报增加，当阈值为 0.25 时，误报和漏报都为 0，达到较好的平衡. 同时，平均延迟随阈值增加而增高. 因此，选择 0.25 作为熵检测的阈值可以降低误报漏报的同时实现较小的检测延迟.

5.3.3　检测窗口比较

根据 Hoeffding 边界确定的窗口大小阈值 1400，并与窗口大小为 1000，1800 时进行对比，验证该边界的可行性，结果如图 5 所示. 从图 5 可以看出，当窗口大小为 1400 时，熵检测方法可以有效地检测到概念漂移点，包括突变和渐变. 从图 5 也可以看出，当窗口越大，检测结果越明显，但带来较大的



(a) WAND 数据集



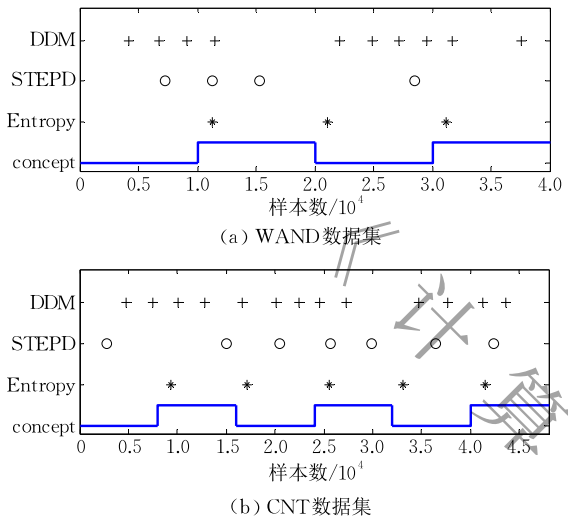
(b) CNT 数据集

图 5　熵检测结果

检测延迟.在图 5(a)中,当窗口大小为 1000 时,在样本 8000 前存在一个误报,因为窗口过小,噪声影响较大.因此,为了尽可能降低检测延迟和噪声影响,根据 Hoeffding 边界将窗口大小阈值设为 1400 是合适的.

5.3.4 检测方法比较

图 6 描述了熵检测方法与两种常用的概念漂移检测方法(DDM,STEPD)的对比结果,DDM[15] 和 STEPD[16] 方法属于基于分类准确性的检测方法.



(a) WAND数据集

(b) CNT数据集

图 6 检测方法比较

从图 6 可以看出,Entropy 检测方法可以有效地检测网络流概念漂移,而 DDM 和 STEPD 检测方法存在较多的误报和漏报,不适用网络流概念漂移检测.DDM 概念漂移检测通过计算分类器的错误率 p_i,以及标准差 $s_i=sqrt(p_i(1-p_i)/i)$,在没有发生概念漂移的情况下 p_i 会随着样本数 i 的增加而呈现总体变小的趋势,最终趋近一个较低的值;但当发生概念漂移时,p_i 和 s_i 会呈上升趋势.实际检测过程中,根据 $p_{\min}$ 和 $s_{\min}$ 判断是否发生概念漂移,当出现 $p_i+s_i<p_{\min}+s_{\min}$ 时,当前的 p_i 和 s_i 分别替换 $p_{\min}$ 和 $s_{\min}$.之后,将概念漂移警告点和确信点定为 $p_i+s_i\geq p_{\min}+2\times s_{\min}$ 和 $p_i+s_i\geq p_{\min}+3\times s_{\min}$,即在 $p_i+s_i\geq p_{\min}+3\times s_{\min}$ 时,判定发生了概念漂移.由于 DDM 只用了分类错误率来检测概念漂移,分类器由于特征子集或模型不稳定,分类性能会有一定的上下浮动,因此容易造成概念漂移检测的误报.另外,由于流量类别的识别率存在差异,占多数的类型识别率相对较高,而少数类型识别率相对较低,因此,不同时间点的类别占比不同,识别准确率也有很大的差异,同样容易产生误判.

检测方法 STEPD 与 DDM 具有相似的架构,该方法采用统计检验来检测概念漂移,通过比较全局准确性和近期准确性来识别概念漂移,该方法假定如果目标概念是不变的,那么分类器对最近 W 个样本的分类准确率将和全局的分类准确率是一致的,如果近期的准确率显著变化则表明发生概念漂移.由于 STEPD 方法只根据识别准确率来检测概念漂移,一方面,由于分类器本身的分类性能不稳定,导致误报的出现;另一方面,对于类不平衡性问题,占大多数的协议概念漂移检测不存在问题,对于占少部分的协议类型,识别准确率变化大,容易造成误报和漏报.另外,DDM 和 STEPD 概念漂移检测方法都需要类别标记,以便统计识别准确率.然而,信息熵检测方法根据特征属性分布的熵变化检测概念漂移,不需要流量的类别信息.

5.4 性能分析

5.4.1 特性选择结果

网络流特征分布会随时间和环境变化产生概念漂移现象,因此,单一特征选择方法在给定数据集获得的特征子集无法在未来长时间维持稳定的分类性能.采用 5 种不同的特征选择方法选取 CNT 和 WAND 数据集的特征子集与 FSEN 方法进行对比,并比较各特征子集建立的朴素贝叶斯模型的分类准确率,分类结果如表 3 所示.

表 3 分类准确率

方法	数据集									
	CNT1	CNT2	CNT3	CNT4	CNT5	WAND1	WAND2	WAND3	WAND4	WAND5
FSEN	96.2	96.5	96.4	95.6	96.5	97.6	99.0	98.7	97.7	97.6
FCBF	92.2	95.0	92.5	92.4	94.5	96.5	94.8	96.6	93.1	91.9
GainRatio	92.7	94.1	90.9	92.8	4.40	91.6	88.8	19.9	89.5	91.4
Chi-square	59.6	95.6	90.0	95.5	95.6	96.9	97.3	97.5	97.5	97.3
InfoGain	44.3	77.5	76.1	94.6	28.4	96.0	96.4	97.8	96.8	97.2
CBC	35.5	13.0	13.5	5.5	18.5	47.0	97.1	90.0	95.3	96.3

表 3 可以看出 FSEN 方法在 CNT 和 WAND 数据集的分类准确率分别高于 95% 和 97%,因为 FSEN 方法通过集成多个特征选择方法获取高泛化能力的特征子集,避免特征选择陷入局部最优.与其他特征选择方法相比,FSEN 方法的准确率较高且稳定,而其他特征选择方法的分类准确率不太稳定,

因为单个特征选择算法考虑的评估指标单一，无法获得全局特征子集，且由于流特征随时间发生变化，之前得到的特征子集无法适应当前的流样本空间，导致分类性能下降. FSEN 特征选择方法获得的特征子集如表 4 所示.

表 4　特征子集

简称	特征描述
Avg_seg_size	平均块大小：字节与包数的商（客户端到服务器端方向）
Init_win_bytes	初始窗口发送的字节数（客户端到服务器端方向及反向）
Data_xmit_time	从第一个包到最后一个非空包的传输时间（客户端到服务器端方向）

FSEN 特征选择后的特征包括 Avg_seg_size，Init_win_bytes，Data_xmit_time. Avg_seg_size 表示平均块大小，不同的应用平均块大小差别很大，如流媒体的平均块大小较大，而即时通信中文本消息的平均块大小相对较小；Init_win_bytes 表示初始窗口发送的字节数，Data_xmit_time 表示流持续时间，可以看出特征子集中的特征区别性都很强，且特征之间也不冗余，表明 FSEN 方法可以集成多个特征选择方法的优势选取高泛化能力的特征子集，有效解决流量分类中的概念漂移问题.

5.4.2　分类器影响因素

本文采用 C4.5 决策树作为基分类器，并且比较 ACED 方法与 WACE 方法在训练数据集大小不同的情况下的分类精度，图 7 描述了初始训练集从 2000～12000 时，两种分类方法性能的差异. 结果显示当初始训练集数目相同时，ACED 方法的分类精度明显高于 WACE 方法. 当初始样本数目为 6000 时，分类准确率最高. 图 8 描述了不同集成分类器数目对分类效果的影响，分类器数目变化从 3～9. 结果显示，初始时增加集成分类器的数目可以提高分类结果的精确度，当初始分类器的数目为 5～7 时，

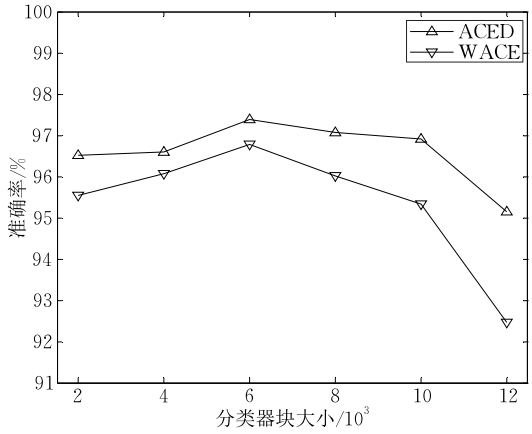


图 7　分类器块大小影响

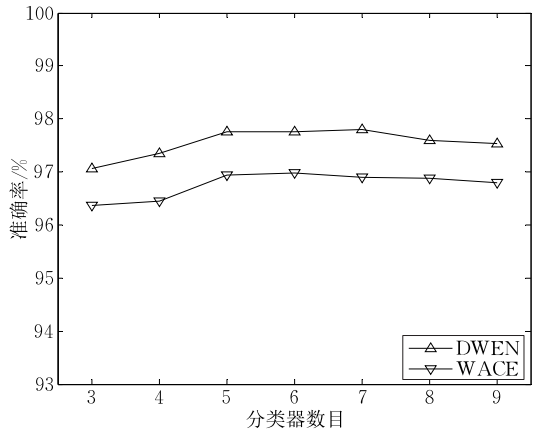


图 8　分类器数目影响

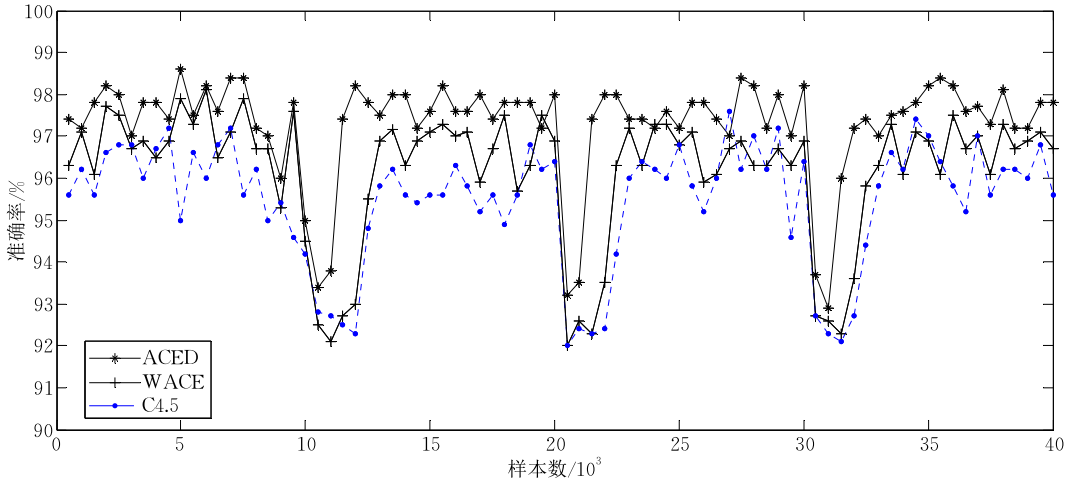
分类效果较好，可以达到 97.7%. 如果再增加集成分类器的数目，反而会降低分类准确率. 因此，将分类器数目定为 5.

5.4.3　分类准确率

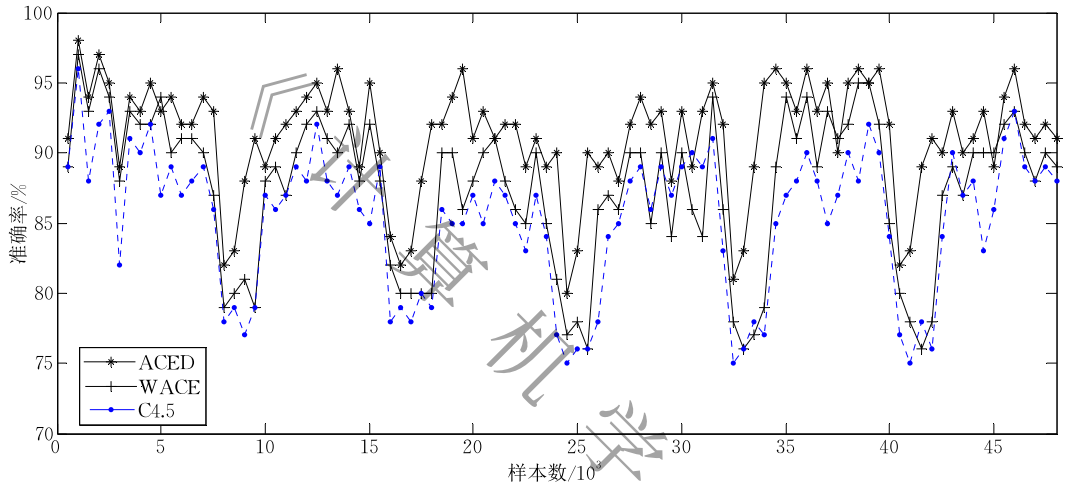
当前网络应用种类繁多，且不断推出新的应用，各应用都具有独特的网络行为特征，但行为特征所表现的统计特征随时间和环境改变会发生概念漂移，之前训练的分类器很难适用于当前样本空间，使得分类性能随之下降，因此，急需自适应网络流概念漂移的分类方法维持较好的分类性能. 本文根据准确率（*Accuracy*）和综合评价（*F*-Measure）评估算法的分类性能. 由于在单分类器中决策树 C4.5 的分类效果较好，所以本文选用 C4.5 作为基分类器，将 3 种算法（ACED、WACE[26] 和 C4.5）进行对比，分类性能如图 9 所示.

从图 9 可见，ACED 算法分类效果明显高于其他分类器. 在检测到概念漂移时 ACED 算法可以快速地更新分类器，延迟明显小于其他算法. 因为 ACED 算法检测到概念漂移再更新分类器，而 WACE 分类器根据固定周期更新. 综合来看，ACED 算法根据检测到的概念偏移点引入新环境的流量重新学习，再更新分类器，可以及时发现概念漂移，降低概念漂移检测延迟带来的影响. 而 WACE 根据固定周期更新分类器，需要耗费更多的时间，且更新效果得不到保证. 而 C4.5 方法重新训练分类器，没有充分利用历史知识.

分类准确率只能综合评价整个数据集的识别精度，算法不仅在整体上要具有较高的分类性能，同时在各个应用上也要具有较高的查全率和查准率，特别当各个应用的样本分布不均匀时，在每个单独应用类别上的识别效果特别重要. *F*-Measure 是综合查准率 *Precision* 和查全率 *Recall* 给出的一个综合评价指标，当 *F*-Measure 较高时则说明方法比较理想，综合评价 *F*-Measure 如图 10 所示.

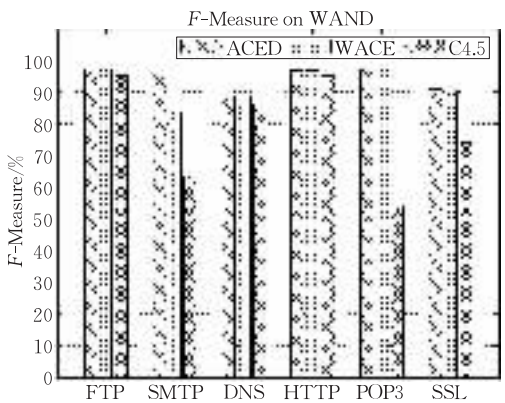


(a) WAND 数据集

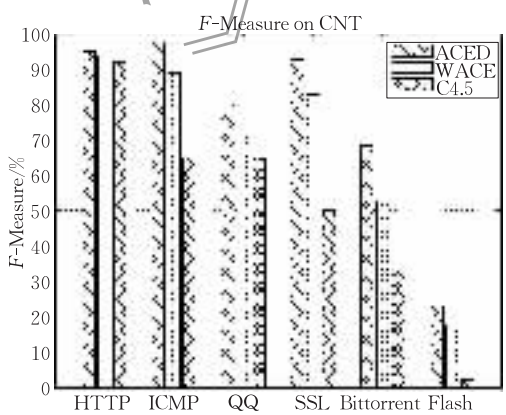


(b) CNT 数据集

图 9 分类准确率



(a) WAND 数据集



(b) CNT 数据集

图 10 *F*-Measure

从图 10 可见,ACED 分类器在单个类别的分类准确率均高于其他分类器. 由于训练样本的类别不平衡,各个类别的样本数目对分类结果有很大影响,样本数目充足,分类的准确率较高;而样本数目稀少,分类效果相对较差. 而 ACED 可以有效提高少数类如 POP3、ICMP 和 Bittorrent 的分类效果.

5.5 分类效率

分类系统的及时反馈可以更好地预判网络异常行为,采取及时准确的应对措施. 为了检验该算法的分类效率,统计 Entropy 算法的检测时间和分类器

更新时间，实验重复 30 次取平均值. 图 11 描述了 ACED 算法的熵检测时间开销，样本集范围分别介于(10 000～80 000)，呈比例增长.

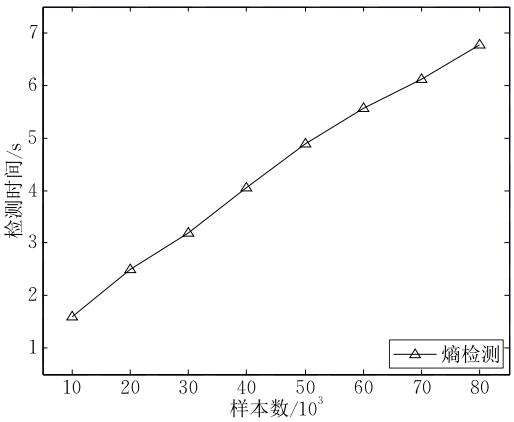


图 11　熵检测时间

从图 11 可以看出，熵检测时间随样本数量增长呈比例增加. 熵检测方法检测 10000 个样本的时间约为 1.6 s. 图 12 描述了不同训练集大小的分类器更新时间开销，训练集范围分别介于(2000～12 000)，呈比例增长.

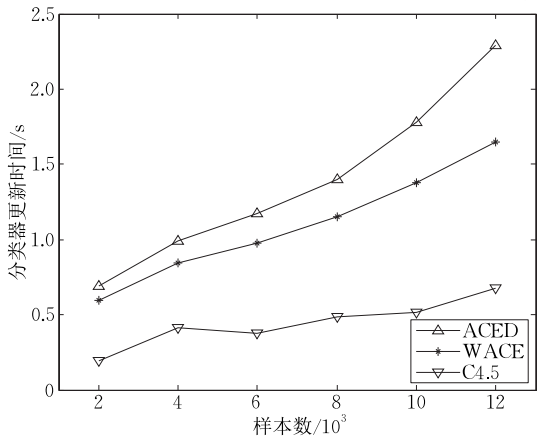


图 12　分类器更新时间

从图 12 可以看出 ACED 和 WACE 方法在分类器更新时间上明显高于 C4.5，因为 ACED 和 WACE 方法是集成方法，更新分类器过程中要建立多个分类器，而 ACED 方法高于 WACE 方法是因为计算得到各分类器的权重后要剔除性能下降的分类器再集成. 为了达到较高的分类时效，可以动态调整样本数量达到较快的分类器更新速度，也可以采用并行计算来提高效率.

6　结　论

随着流量产生环境的变化，流特征和分布随之发生改变，概念漂移的发生使得分类器很难维持较高的分类性能. 本文提出一种基于信息熵的自适应网络流概念漂移分类方法，根据特征属性的信息熵变化检测概念漂移，并根据增量集成学习策略引入新样本建立的分类器，并替换性能下降的分类器，最后，将加权集成分类器更新概念漂移前的分类器. 实验结果表明该方法可以有效检测概念漂移，并学得自适应分类器，有效应对流量分类中存在的概念漂移问题. 下一步工作将结合代价敏感学习解决概念漂移中的类别不平衡问题.

致　谢　审稿人对论文提出了宝贵意见，在此表示感谢！

参　考　文　献

[1] Finsterbusch M, Richter C, Rocha E, et al. A survey of payload-based traffic classification approaches. Communications Surveys & Tutorials, 2014, 16(2): 1135-1156

[2] Dainotti A, Pescape A, Claffy K C. Issues and future directions in traffic classification. Network, 2012, 26(1): 35-40

[3] Grimaudo L, Mellia M, Baralis E, et al. Self-learning classifier for Internet traffic//Proceedings of the IEEE INFOCOM 2013. Turin, Italy, 2013: 3381-3386

[4] Jin Y, Duffield N, Erman J, et al. A modular machine learning system for flow-level traffic classification in large networks. ACM Transactions on Knowledge Discovery from Data, 2012, 6(1): 4

[5] Lee S, Kim H, Barman D, et al. Netramark: A network traffic classification benchmark. ACM SIGCOMM Computer Communication Review, 2011, 41(1): 22-30

[6] Zhang H, Lu G, Qassrawi M T, et al. Feature selection for optimizing traffic classification. Computer Communications, 2012, 35(12): 1457-1471

[7] Raahemi B, Zhong W, Liu J. Peer-to-peer traffic identification by mining IP layer data streams using concept-adapting very fast decision tree//Proceedings of the 2008 20th IEEE International Conference on Tools with Artificial Intelligence. Dayton, USA, 2008, 1: 525-532

[8] Liu Q, Liu Z, Wang R, et al. Large traffic flows classification method//Proceedings of the 2014 IEEE International Conference on Communications Workshops (ICC). Sydney, Australia, 2014: 569-574

[9] Schlimmer J C, Granger R H. Beyond incremental processing: Tracking concept drift//Proceedings of the 5th National Conference on Artificial Intelligence. Philadelphia, Pennsylvania, 1986: 502-507

[10] Gama J, Žliobaitė I, Bifet A, et al. A survey on concept drift adaptation. ACM Computing Surveys, 2014, 46(4): 44

[11] Zhong W, Raahemi B, Liu J. Classifying peer-to-peer applications using imbalanced concept-adapting very fast decision tree on IP data stream. Peer-to-Peer Networking and Applications, 2013, 6(3): 233-246

[12] Hulten G, Spencer L, Domingos P. Mining time-changing data streams//Proceedings of the 7th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Francisco, USA, 2001: 97-106

[13] Erman J, Mahanti A, Arlitt M, et al. Semi-supervised network traffic classification//Proceedings of the 2007 ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Systems. San Diego, USA, 2007, 35(1): 369-370

[14] Raahemi B, Zhong W, Liu J. Exploiting unlabeled data to improve peer-to-peer traffic classification using incremental tri-training method. Peer-to-Peer Networking and Applications, 2009, 2(2): 87-97

[15] Gama J, Medas P, Castillo G, et al. Learning with drift detection//Proceedings of the 20th Brazilian Symposium on Artificial Intelligence Artificial Intelligence. Sao Luis, Brazil, 2004: 66-112

[16] Nishida K, Yamauchi K. Detecting concept drift using statistical testing//Proceedings of the 10th International Conference Discovery Science. Sendai, Japan, 2007: 264-269

[17] Bifet A, Gavalda R. Learning from time-changing data with adaptive windowing//Proceedings of the SIAM International Conference on Data Mining. Minneapolis, Minnesota, 2007: 443-448

[18] Du L, Song Q, Jia X. Detecting concept drift: An information entropy based method using an adaptive sliding window. Intelligent Data Analysis, 2014, 18(3): 337-364

[19] Lu N, Zhang G, Lu J. Concept drift detection via competence models. Artificial Intelligence, 2014, 209: 11-28

[20] Gringoli F, Salgarelli L, Dusi M, et al. GT: Picking up the truth from the ground for internet traffic. ACM SIGCOMM Computer Communication Review, 2009, 39(5): 12-18

[21] Jin R, Agrawal G. Efficient decision tree construction on streaming data//Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Washington, USA, 2003: 571-576

[22] Kirkby R, Bouckaert R R, Studen M, et al. Improving Hoeffding trees. International Journal of Approximate Reasoning, 2007, 45: 39-48

[23] Este A, Gringoli F, Salgarelli L. On the stability of the information carried by traffic flow features at the packet level. ACM SIGCOMM Computer Communication Review, 2009, 39(3): 13-18

[24] Pan Wu-Bin, Cheng Guang, Guo Xiao-Jun, et al. An embedded feature selection using selective ensemble for network traffic. Chinese Journal of Computers, 2014, 37(10): 2128-2138(in Chinese)
(潘吴斌, 程光, 郭晓军等. 基于选择性集成策略的嵌入式网络流特征选择. 计算机学报, 2014, 37(10): 2128-2138)

[25] Bernaille L, Teixeira R, Akodkenou I, et al. Traffic classification on the fly. ACM SIGCOMM Computer Communication Review, 2006, 36(2): 23-26

[26] Wang H, Fan W, Yu P S, et al. Mining concept-drifting data streams using ensemble classifiers//Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Washington, USA, 2003: 226-235

PAN Wu-Bin, born in 1987, Ph. D. candidate. His research interests include network security, network measurement and traffic classification.

CHENG Guang, born in 1973, Ph. D., professor, Ph. D. supervisor. His research interests include network security, network measurement and behavior, future Internet security.

GUO Xiao-Jun, born in 1983, Ph. D. candidate. His research interests include network security, network measurement and network management.

HUANG Shun-Xiang, born in 1991, M. S. candidate. His research interests include network security, network measurement, and traffic classification.

Background

Network traffic classification is a basic technology in network management and security. Machine learning algorithms are widely used to recent traffic classification, and feature selection is used as a pre-processing step to eliminate irrelevant and redundant features, but existing feature selection techniques metrics single, and overlook the intrinsic link of characteristic properties and application itself, and sensitive to variation in traffic data, leading it difficult to keep a better performance. Meanwhile, class imbalance and concept drift affect feature selection to obtain stable feature subset. Therefore, an embedded feature selection based on selective ensemble is proposed, according to the selective ensemble strategy to ensemble part of feature selectors to get a better performance than ensemble all. And then through

the combination method of improved sequence forward search and wrapper secondary search optimal feature subset. Finally, the subset with highest accuracy is selected as a global feature subset. The proposed algorithm can eliminate irrelevant and redundant features, and improve the stability of feature subset effectively.

Moreover, a metric of stability is introduced to measure the feature subset, the metric of weighted occurrence frequency is used to evaluate whole feature subset, and it can deal with concept drift effectively. Experimental results show that the proposed algorithm can reduce the complexity of feature subset effectively while ensuring the classification performance, so as to achieve the optimal balance of the classification performance, efficiency and stability.

This work was supported by the National High Technology Research and Development Program (863 Program) of China (2015AA015603), the Prospective Research Programs Future Internet of Jiangsu Province (BY2013095-5-03), the Six Talent Peaks of High Level Talents Project of Jiangsu Province (2011-DZ024), the Fundamental Research Funds for the Central Universities and the Research and Innovation Project for College Graduates of Jiangsu Province (KYLX15_0118).

The research team has focused on network management and security for years, and more than ten papers in this domain have published in highly-ranked conferences and journals.