基于 M 值概率分布的网络视频流分类

杨凌云[①②] 董育宁[*①] 王再见[②] 汤萍萍[①②]
[①](南京邮电大学通信与信息工程学院 南京 210003)
[②](安徽师范大学物理与电子信息学院 芜湖 241000)

摘 要：为了改善网络视频流的细粒度分类效果，该文分析视频流传输过程中的特征变化与流分类之间的关系。根据不同类型的视频流具有不同的下行传输速率变化模式，提出一种新的基于下行速率传输的视频流分类特征 — M 值概率分布，并使用支持向量机(SVM)实现网络视频流的分类。实验结果表明，M 值概率分布相比较于已有的常见流特征，可以更好地实现 6 种典型的网络视频流分类。

关键词：网络视频流；流分类；M 值概率分布
中图分类号： TP393 **文献标识码**： A **文章编号**： 1009-5896(2018)05-1094-07
DOI: 10.11999/JEIT170617

Network Video Traffic Classification Based on Probability Distribution of M Value

YANG Lingyun[①②] DONG Yuning[①] WANG Zaijian[②] TANG Pingping[①②]
[①](*College of Communication and Information Engineering*, *Nanjing University of Posts and Telecommunications*, *Nanjing* 210003, *China*)
[②](*College of Physics and Electronic Information*, *Anhui Normal University*, *Wuhu* 241000, *China*)

Abstract: To obtain better results for fine-grained video traffic classification, this paper analyzes the relationship between the feature variations during transmission and video traffic classification. According to the nature that different types of video services contain different downlink transmission rate variation patterns, a new video flow feature — M value probability distribution, based on downlink byte rate variation is proposed, and video classification is realized by Support Vector Machine (SVM). The experimental results show that the probability distribution of M value is a better feature for classification of six kinds of common network video applications than other commonly used flow features.

Key words: Network video; Traffic classification; Probability distribution of M value

1 引言

随着信息和通信技术的迅猛发展，由此产生的数据及增长速度都处于高速增长阶段，人们进入了‘大数据’ 的时代。同时，随着互联网技术的高速发展，多媒体数据信息流特别是视频流信息在网络通信中所占的比重越来越大，如何对这些多媒体数据流进行分析和识别是当前研究的一个热门话题[1]。运用机器学习(machine learning)算法挖掘数据有用信息成为近年来人们研究的重点[2]。机器学习算法分为监督学习算法[3–5]、非监督学习算法[6,7]和半监督学习算法[8,9]。监督学习算法因为相比较另外两种算法具有较好的分类效果而得到广泛认可和应用。

近年来，多媒体数据特别是视频流几乎无处不在，由于视频流数据包含丰富的信息，如何从视频流中提取有效信息，面临着巨大挑战[10]。文献[11]中提出了 248 种常见的流特征属性，包括包大小(均值、方差)，包达到时间差等很多目前依然常用的特征。文献[12,13]采集了 20 个统计特征中讨论特征选取方法及分类。文献[14]在上面 20 个特征的基础上加入了主机响应持续时间，形成 21 个基础特征，具有代表性。文献[15]加入了协议特征，TCP 协议、UDP 协议的比例也成为流分类的特征之一。不同类型的网络流对特征的要求也是不同，文献[16]对基础数据、服务、P2P, FTP、邮件、HTTP 等进行分类，选取了 29 种特征进行分析讨论。文献[17]对音乐、视频、邮件等 12 种不同应用分类。文献[18]针对视频流选取上下行速率作为分类的主要特征。文献[19]

收稿日期：2017-06-28；改回日期：2018-02-23；网络出版：2018-03-16
*通信作者：董育宁 dongyn@njupt.edu.cn
基金项目：国家自然科学基金(61271233, 61401004, 61601005)，华为 HIRP 创新项目，安徽师范大学博士科研启动金项目(2016XJJ129)
Foundation Items: The National Natural Science Foundation of China (61271233, 61401004, 61601005), The HIRP Program of Huawei Technology Co. Ltd, The Ph.D Programs Foundation of Anhui Normal University (2016XJJ129)

对微信流进行细粒度分类，文献[20]提出新的特征——平均字节峰值，实现视频流 3 种不同清晰度的分类。但这些流统计特征并没有考虑流在传送过程中的特征变化，特别是在视频流中，不同类型的传输流的传输过程是不同的，本文的创新点就是根据流传输过程中的数据变化提取新的流分类特征。

本文研究发现不同类型的视频流的传输速率的变化是不同的，创新点就是找出典型视频流的内部变化特征，通过观察标准化后的下行速率变化，计算下行速率的离散概率分布值(即 M 值概率分布)，并把它作为一组新的分类特征。实验结果表明，这些新特征可以较好地实现 6 种典型的视频流分类。M 值概率分布指的是以下行速率为基础，它首先把一条视频流分成若干个小的时间片，分别计算出不同时间片的平均下行速率值，通过量化下行速率值求得各个时间片的 M 值，然后求出每条流的 M 值概率分布作为视频流分类特征值。

特征选择算法指的是从众多原始特征中选择适合网络业务分类的最佳特征子集，典型的特征选择算法包括滤波式(filter)、封装式(wrapper)和嵌入式(embedded)[21]，滤波式算法通常与分类器的选择无关，主要包括 CFS(Correlation based Feature Selection)[22], ReliefF[23]及增益比(gain ratio)等。本文在特征选择中结合 CFS 和 ReliefF 算法，CFS 算法是基于相关性的一种特征子集算法，但因为没有考虑冗余问题，只用 CFS 无法达到子集最优，ReliefF 考虑了同类特征与异类特征综合因素进行特征选择。研究发现，两者结合更能实现特征的最优子集选择。最后使用 SVM 和 logistic 分类器实现网络视频流的分类。

论文具体安排如下：第 1 节为引言，主要介绍现有的网络流的分类特征和分类方法研究的主要特点。第 2 节详细叙述新提出的 M 值量化概率分布，它以下行速率为基础，通过分片，量化，求概率分布得到离散的概率分布值，然后用支持向量机(SVM)分类方法实现网络视频流分类。第 3 节为实验验证部分，主要描述 3 种情况下的不同类型的视频流的细粒度分类结果，讨论分片大小等参数对分类结果的影响。最后是结论。

2　M 值概率分布

目前，对视频流业务分类特征的提取主要集中在整体数据结果的统计上，不同类型的视频流的这些数据特征参数有相似的内容，也有不同的，但是不同类型视频流的的数据传输过程都具有自己的特征，与其他视频流存在明显不同。概率分布是讨论数据传输过程的特征参数，但是对于概率分布的统计通常体现在信息熵的度量(包大小信息熵，包时间间隔熵等)，但这些是远远不够的，如何在实际应用中利用数据传输过程的不同，通过概率分布特征实现不同类型的网络视频流分类，是本文的创新点。M 值概率分布是根据下行速率的传输变化的不同作为视频流的分类特征，量化视频流传输过程中的速率的概率分布值，进而提取分类特征，实现网络视频流的分类。

图 1 共给出了 6 种不同的视频流(优酷标清 [1]、优酷高清、优酷超清、交互式视频(以 QQ 为例)[2]、直播式视频(以 cbox[3], sopcast[4]为例)、P2P 非直播视频(以迅雷看看为例)[5])在 15 min 内的传输过程中的下行速率的变化情况，每 1000 个数据包计算一次下行速率均。由图 1 可以看出，非对称的视频流经常会出现一段时间速率很快，一段时间速率又很低的情况，那是因为数据通常是先缓冲到用户端的缓冲器中，然后再播放视频，由于视频缓冲器中数据较多，通常会暂时减少数据的发送，直到缓冲器中的数据减少到一定的值，会再次集中填充缓冲器。同时，对称视频流比非对称速率的变化更加缓和，因为对称式视频相对于非对称式视频的缓冲相对较少，特别是对交互式视频和直播式视频。

M 值概率分布指的是针对视频流的下行速率在整个流传输过程中的变化，求取离散概率分布的过程，具体步骤如下：

步骤 1　对视频流进行分块。

数据流的分块的大小会直接影响到分类效果，选取的数据块过大，会掩盖掉下行速率的变化过程，从而对后面的分类带来误差，所以选取的数据块应尽可能小些，但数据太小会模糊了一些变化的规律，从而不能达到最优分类的目的。

步骤 2　运用式(1)对每个数据块求取下行速率。

$$V_{\mathrm{BR}} = \frac{B_S \cdot 8}{t_S \cdot 1024} \tag{1}$$

其中，B_S 为一定时间内下行传输的总的字节数，t_S 为传输结束的时间减去开始的时间，单位为 s，下行传输速率 V_{BR} 的单位为 kbps(千比特率)。

步骤 3　运用式(2)对每个数据块的下行字节速率进行归一化。

$$M = 1 - \frac{1}{K + V_{\mathrm{NBR}}^{K_1}} \tag{2}$$

[1] http://www.youku.com

[2] http://im.qq.com/

[3] http://cbox.cntv.cn/

[4] http://www.sopcast.com

[5] http://www.kankan.com/

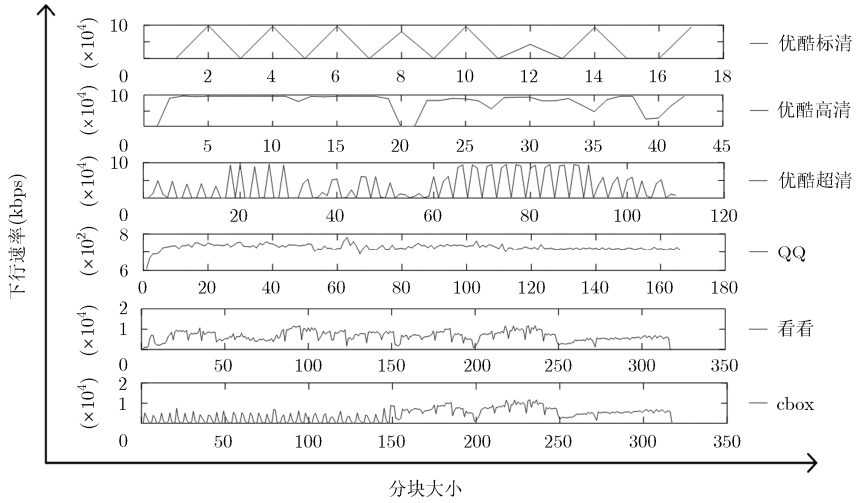

图 1 6 种典型视频流的下行速率变化过程

M 值范围为$(1-1/K,1)$，其中 K 是一个常量，K 值的变化会扩大或者缩小量化范围，同时也会影响到分类的结果，K_1的主要作用也是为了区分低速率值，一般设置为 1。V_{NBR} 为标准化速率，它的计算公式为[24]

$$V_{\mathrm{NBR}}=\frac{8\times 30\,V_{\mathrm{BR}}}{1000\min(30,\mathrm{FR})} \tag{3}$$

其中，FR 为视频流的帧速率，V_{BR} 为下行速率。

归一化 M 值的计算优点主要体现在两个方面：(1)如果不对下行速率进行归一化，那么我们得到的速率的值变化跨度是非常大的，最大值和最小值之间有上千倍的关系，从这些相差比较大的数值中，选取合适的值进行速率的离散概率分布求取，几乎是不可能的。(2)M 值的计算是具有可扩展性的，既可以实现视频流的分类，也可以实现较低速率。

步骤 4 求取 M 值离散概率分布。

通过步骤 3 的归一化，M 值集中在一定的范围内，对 M 值进一步量化，在其范围内求取 5 个概率分布特征值。例如：$k=1$，那么 M 的范围就在 0~1 之间，根据视频流特征的 M 值分布，求取 M 值的 5 个离散概率分布值。图 2 共给出了 6 种不同的视频流传输过程中的 M 值概率分布值。通过图 2 可以明显地看到它们之间各自的特点，所以 M 值概率分布可以作为特征用于不同类型的视频流的分类。

M 值概率分布以下行字节速率为基础，讨论整个流传输过程的变化特性，从而实现不同类型的视频类型的分类，扩展之，不同类型的网络应用都有自己的内部传输属性，以此作为特征，实现网络流分类似乎是可行的。

3 实验结果

实验采取的数据是分别在笔记本和台式电脑上

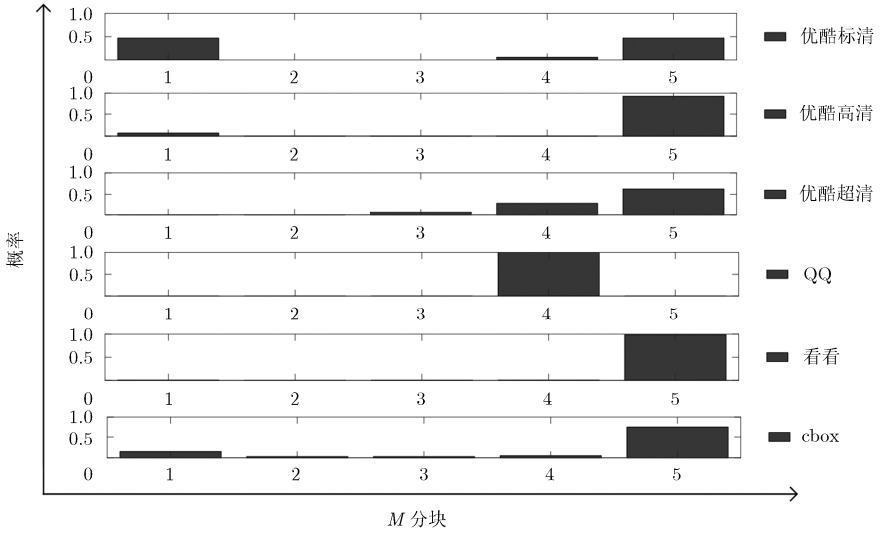

图 2 6 种不同清晰度视频流的 M 值概率分布

使用 Wireshark[6)]通过南京邮电大学校园网采集获得，网关 IP 地址为 10.10.143.20 和 10.10.145.20，主机 IP 地址主要有 10.10.145.47, 10.10.143.47, 10.10.143.7，获得数据的主机分别位于南京邮电大学科技楼和有线楼，数据的获取是有线和无线相结合的方式。采集的视频有：优酷 3 种不同清晰度的视频流(标清，高清和超清)，P2P 文件共享类视频(迅雷看看)，交互式视频(QQ)，网络直播式视频(cbox, sopcast)。每条视频流的长度为 15 min 左右(参见表 1)。通过 Windows 批处理软件处理(.bat)后保留五元组信息，数据处理及特征提取采用的是 MATLAB (R2016a)软件，分类器以 SVM 为基础进行分类精度测试，同时结合线性 SVM 和逻辑回归分类算法，对典型视频流进行分类。分类精度测试采用 2 折交叉验证的方式。

表 1 网络视频数据集

业务类型名	样本大小(GB)	样本数(900 s)
优酷标清	43.86	119
优酷高清	68.53	120
优酷超清	146.70	119
cbox,sopcast	218.63	120
迅雷看看	317.56	120
QQ 视频	76.24	120

在使用分类器进行分类时，数据存在着正确分类和误判的情况，衡量网络流分类的好坏通常用正确率 A(Accuracy)、召回率 R(Recall)、精确度 P(Precision)、F-值 F(F-measure)。

3.1 非对称式视频流分类

非对称式视频流分类采用典型的优酷的 3 种不同清晰度的视频流，实现分类，即优酷标清、高清和超清视频，量化速率的概率分布可以实现这 3 种不同视频的细粒度分类。

通过表 2 可以看出，随着 K 值及分块大小的变化，M 值概率分布作为特征实现 3 种不同清晰度的视频流的结果是不断变化的；同时也可以看出 K 值和数据包的大小对分类结果有影响。选择 k 值和数据包的大小是基于以下几个基本原则：(1)数据包的分块个数在 1000～3000 个数据包之间，既能反映速率特征的变化过程，又滤除了噪声的干扰。(2)K 值在数据包个数确定后取分类精度最高的那个。(3)如果 1000～3000 个数据包的分类精度在 2%误差范围内，选择数据包较小的那个。(4)尽量保证所选参数具有普遍性，也就是说比较容易得到，就算是改变数据包个数或者 K 值，变化相对稳定。所以实验选择的数据包大小为 1000 数据包，K 等于 2。

表 2 非对称式视频流分类中数据分块大小及 K 值变化对分类正确率的影响(%)

K 值	分块(包数目)						
	100	300	500	1000	2000	3000	5000
1.0	83.0	88.3	90.8	93.0	94.6	93.6	94.4
1.5	82.1	87.4	91.6	92.5	95.5	94.7	91.3
2.0	83.0	90.8	93.9	94.1	95.0	91.9	92.2
2.5	84.1	92.2	92.5	94.1	93.0	93.0	93.5

文献[12]给出的 20 个关于网络流分类的特征，是在采集到视频流后，通常比较容易得到，而且相对有效的特征，极具代表性；同时文献[12]采用的 CFS 算法是基于相关性的特征选择算法，非常具有普遍性，所以本文以文献[12]的方法作为对比。根据上述原则，M 值概率分布值选择 1000 个数据包分块，$k=2$ 时的结果。初始特征为文献[12]的特征加上本文提出的 5 个 M 值概率分布特征(按照量化顺序从小到大分别命名为 $M1$～$M5$)及下行速率，共 26 个特征。

由表 3 可以看出，运用 ReliefF 算法对非对称视频流的 26 个初始特征进行排序结果，说明 M 值概率分布特征可以较好地实现非对称视频流 3 种不同清晰度的分类。

表 3 非对称视频流-ReliefF 算法对特征的排序(前 10 个特征)

序列号(排序)	特征名称	特征权值
1	$M5$	0.21171
2	$M3$	0.20563
3	下行字节数	0.16401
4	下行速率	0.16366
5	下行包个数	0.11882
6	$M4$	0.10861
7	下行包方差	0.08488
8	上行包大小	0.06315
9	$M2$	0.04305
10	下行包均值	0.03636

表 4 给出了本文方法与文献[12]中使用 CFS 特征选取方法的分类在查准率、召回率和 F-测度结果的对比，文献[12]中 CFS 选取上行字节数，上行最大包大小，下行字节数，下行包均值 4 个特征，本文选取下行字节数，$M3$ 和 $M5$ 共 3 个特征，除了超清视频的查准率文献[12]中的结果等于本文方法(都为 98.3%)外，其它类型的视频流的 3 个主要评价参数(查准率，召回率，F-测度)都高于文献[12]中的方法。

6) https://www.wireshark.org/

表 4 非对称视频流分类结果对比(SVM 分类器)

非对称视频流		查准率	召回率	F-测度
标清	文献[12]方法	0.983	0.966	0.975
	本文方法	0.983	0.983	0.983
高清	文献[12]方法	0.902	0.925	0.914
	本文方法	0.935	0.958	0.947
超清	文献[12]方法	0.924	0.916	0.920
	本文方法	0.974	0.950	0.962

表 5 给出了 Weka[7)]中的 3 种不同的分类器在两种分类特征中的分类精度值对比，这 3 种分类器分别是常规 SVM 分类器(libsvm)，线性 SVM 分类器(liblinear)，基于逻辑回归的岭估计分类器(logistic)，从分类结果可以看出，本文方法在这 3 种分类器中，精度值都有提高，其中 liblinear 方法和 logistic 方法精度值均为 97.8%。

表 5 非对称视频流的分类精度对比(不同分类器)(%)

方法	分类器		
	libsvm	liblinear	logistic
文献[12]方法	93.6	92.1	90.6
本文方法	96.3	97.8	97.8

3.2 对称式视频流分类

对称式视频流主要是指接收和发送方的地位是平等的，文献[7,13,16]是关于 P2P 视频流分类，也属于对称式视频流。常见的 P2P 应用包括交互式视频，直播式视频和非直播式视频，分别以 QQ 视频，cbox, sopcast 网络直播和迅雷看看实现 P2P 非直播视频为例。

表 6 是视频流分块大小及 K 值在 3 种典型的视频流分类中的影响，对称式视频流的分类效果在分块数值较小时，正确率很高，几乎可以达到 100%，对称式视频流中数据缓冲内容较少，在数据较少时的速率变化规律已经形成，但数据分块较大时，几乎完全掩盖了这些变化规律，看不出速率的变化，分类精度下降较快。同时可以看出 K 值对分类的影响相对较小。K 值及数据分块的选择保持不变。

表 6 对称式视频流数据分块大小及 K 值变化对分类精度的影响(%)

K值	分块						
	100	300	500	1000	2000	3000	5000
1.0	100	100	100	99.7	100	100	91.9
1.5	100	100	100	100	99.7	99.7	91.9
2.0	100	100	99.7	100	99.7	100	91.7
2.5	100	100	99.7	99.7	99.7	100	79.7

7) http://weka.wikispaces.com/

由表 7 可以看出，在对称式视频流分类特征的 ReliefF 权值排序中，M 值概率分布的 5 个特征值比较靠前，说明 M 值概率分布特征可以较好地适用于对称视频业务的分类。

表 7 对称式视频流-ReliefF 算法对特征的排序(前 10 个特征)

序列号(排序)	特征名称	特征权值
1	$M5$	0.432006
2	$M3$	0.414957
3	下行包方差	0.214654
4	下行包均值	0.198844
5	$M2$	0.183646
6	上行字节数	0.173803
7	$M4$	0.168303
8	下行字节数	0.092040
9	下行速率	0.092025
10	下行包个数	0.082754

表 8 可以看出，本文方法无论是查准率、召回率还是 F-测度都达到了 100%，同时相比较于文献[12]的 CFS 算法，本文的方法实现 100%分类的特征数只有 3 个(分别是上行字节数，$M5$ 和下行速率)，而文献[12]的特征数为 11 个，所以本文方法既提高了正确率，又显著减少了特征的个数。

表 8 对称式视频流分类对比(SVM 分类器)

对称式视频流		查准率	召回率	F-测度
直播式	文献[12]方法	1.000	0.975	0.987
	本文方法	1.000	1.000	1.000
非直播式	文献[12]方法	0.976	1.000	0.988
	本文方法	1.000	1.000	1.000
交互式	文献[12]方法	1.000	1.000	1.000
	本文方法	1.000	1.000	1.000

表 9 同样给出了 3 种不同的分类器对比结果，可以看出，无论是在文献[12]中的方法还是本文方法，都高于 99%，提高并不是很明显，本文方法的主要优势就是特征个数的减少。

3.3 网络视频流的分类

本节把前面的所有情况的网络视频流分类，即直播式视频，交互式视频，P2P 非直播式，优酷标清视频，优酷高清视频和优酷超清视频。

表 9 对称视频流的分类精度对比(不同分类器)(%)

方法	分类器		
	libsvm	liblinear	logistic
文献[12]方法	99.2	99.2	99.2
本文方法	100.0	100.0	99.4

表 10 中给出了 6 种常见网络视频流的分类中 K 值和数据分块的影响。同时本文视频流特征与其他特征相结合，可以实现较高视频流分类准确率。

表 10　6 种典型网络视频流数据分块大小及 K 值变化对分类精度的影响(%)

K值	分块						
	100	300	500	1000	2000	3000	5000
1.0	81.5	85.7	88.2	88.0	91.4	90.9	82.6
1.5	82.6	86.5	87.2	90.3	89.7	88.7	83.0
2.0	83.6	87.7	88.0	87.0	88.9	86.6	78.6
2.5	83.0	88.9	87.7	89.5	90.9	90.0	76.5

表 11 中 ReliefF 特征排序结果表明 M 值概率分布的权值也基本高于其他特征，M 值概率分布特征可以较好地适用于 6 种网络视频业务的分类。

表 11　6 种典型网络视频流-ReliefF 算法对特征的排序(前 10 个特征)

序列号(排序)	特征名称	特征权值
1	下行包均值	0.296668
2	$M5$	0.292358
3	$M3$	0.225818
4	下行包方差	0.216628
5	$M4$	0.174624
6	上行字节数	0.133731
7	$M2$	0.108964
8	上行包均值	0.080091
9	下行包个数	0.079283
10	下行字节数	0.079253

表 12 的网络视频流分类中，给出了 6 种典型视频流的分类结果(网络直播视频，非直播视频，交互视频，超清视频，高清视频和标清视频)，可以看出，本文方法优于文献[12]中的方法，在特征选择方面文献[12]的方法选择 10 个特征，本文方法选择了 8 个特征，同时实现了既减少了特征个数，又改善了分类效果。

表 13 给出了 3 种分类器的分类结果。可以看出，本文方法正确率都高于文献[12]中的方法，精度最高的是 liblinear 分类器，精度值是 98.6%。

通过上述 3 个小节的实验，可以看出，因为视频流的传输具有相关性，对长视频流进行分类时，考虑流在传输过程中的变化，并提取出相关特征，不仅可以提高视频流的分类精度，也可以降低特征维数。事实上，考虑整个视频流在传输过程的变化时，不仅已经把某些已有的相关特征包含进去，同时也具备了之前特征集不具有的信息特征，才实现既提高分类精度，又减少了分类特征数的效果。

表 12　6 种典型网络视频流分类结果(SVM 分类器)

	视频类型	查准率	召回率	F-测度
直播式	文献[12]方法	1.000	0.983	0.992
	本文方法	1.000	1.000	1.000
非直播式	文献[12]方法	0.869	0.992	0.926
	本文方法	0.968	1.000	0.984
交互式	文献[12]方法	1.000	1.000	1.000
	本文方法	1.000	1.000	1.000
标清	文献[12]方法	0.973	0.924	0.948
	本文方法	0.983	0.975	0.979
高清	文献[12]方法	0.921	0.875	0.897
	本文方法	0.950	0.950	0.950
超清	文献[12]方法	0.897	0.874	0.885
	本文方法	0.966	0.941	0.953

表 13　6 种典型网络视频流分类精度对比(不同分类器)(%)

方法	分类器		
	libsvm	liblinear	logistic
文献[12]方法	94.1	94.6	92.1
本文方法	97.8	98.6	96.6

4　结束语

M 值概率分布通过研究数据传输过程中的特点，以下行速率为基础，提取用来实现视频流分类的特征，较好地实现了网络视频流的分类，提高了现有特征的分类精度。接下来也可以探讨其他特征参数或者多种特征的组合，讨论特征的过程属性作为分类特征，实现网络流的分类。同时，基于最优数据包分块和 K 值的取值会对最终分类结果产生重要的影响，接下来可以利用深度学习求取相应最优化的值，在已有概率分布的基础上进行特征选择，进一步优化分类结果。

本文关于概率分布特征提取只讨论了视频流的分类情况，也可以扩展到其它的多媒体应用分类或者其它较长时间的流分类中(例如不同游戏流的分类等)。但提取概率分布作为流分类特征的缺点就是需要较长时间的数据采集，而且采集时间越长，数据流的传输特性越明显，流分类的效果也就越好，但是如果数据流时间过短，则无法实现较好的分类结果。

参 考 文 献

[1] ANDERSSON R. Classification of video traffic: An evaluation of video traffic classification using random forests and gradient boosted trees[D]. [Master dissertation], Karlstad University, 2017.

[2] KESAVARAJ G and SUKUMARAN S. A study on classification techniques in data mining[C]. Proceedings of the 4th International Conference on Computing, Communications and Networking Technologies,

Tiruchengode, India, 2014: 1–7. doi: 10.1109/ICCCNT.2013.6726842.

[3] GHOFRANI F, JAMSHIDI A, and KESHAVARZ-HADDAD A. Internet traffic classification using Hidden Naive Bayes model[C]. Proceedings of the 23rd Iranian Conference on Electrical Engineering, Tehran, Iran, 2015: 235–240. doi: 10.1109/IranianCEE.2015.7146216.

[4] MUNTHER A, ALALOUSI A, NIZAM S, *et al.* Network traffic classification — A comparative study of two common decision tree methods: C4.5 and Random forest[C]. Proceedings of the 2nd International Conference on Electronic Design, Penang, Malaysia, 2014: 210–214. doi: 10.1109/ICED.2014.7015800.

[5] HAO Shengnan, HU Jing, LIU Songyin, *et al.* Improved SVM method for internet traffic classification based on feature weight learning[C]. Proceedings of the Fourth International Conference on Control, Automation and Information Sciences (ICCAIS) Changshu, China, 2015: 102–106. doi: 10.1109/ICCAIS.2015.7338641.

[6] VINUSHREE N, HEMALATHA B, and KALIAPPAN V K. Efficient kernel-based fuzzy C-means clustering for pest detection and classification[C]. Proceedings of the 2014 Computing and Communication Technologies (WCCCT), Tamilnadu, India, 2014: 179–181. doi: 10.1109/WCCCT.2014.61.

[7] ZHANG Shichao, LI Xuelong, ZONG Ming, *et al.* Efficient kNN classification with different numbers of nearest neighbors[J]. *IEEE Transactions on Neural Networks and Learning Systems*, 2017: 1–12. doi: 10.1109/TNNLS.2017.2673241.

[8] WANG Pu, LIN Shihchun, and LUO Min. A framework for QoS-aware traffic classification using semi-supervised machine learning in SDNs[C]. Proceedings of the 13th IEEE International Conference on Services Computing, San Francisco, USA, 2016: 760–765. doi: 10.1109/SCC.2016.133.

[9] GLENNAN T, LECKIE C, and ERFANI S M. Improved classification of known and unknown network traffic flows using semi-supervised machine learning[C]. Proceedings of the Australasian Conference on Information Security and Privacy, QLD, Australia, 2016: 493–501. doi: 10.1007/978-3-319-40367-0-33.

[10] BAGHERZADEH-KHIAVANI F, RAMEZANKHANI A, AZIZI F, *et al.* A tutorial on variable selection for clinical prediction models: Feature selection methods in data mining could improve the results[J]. *Journal of Clinical Epidemiology*, 2016(71): 76–85. doi: 10.1016/j.jclinepi.2015.10.002.

[11] MOORE A, ZUEV D, and CROGAN M. Discriminators for use in flow-based classification[R]. Queen Mary University of London, 2013: 1–14.

[12] ZhANG JUN, YANG XIANG, WANG YU, *et al.* Network traffic classification using correlation information[J]. *IEEE Transactions on Parallel and Distributed Systems*, 2013, 24(1): 104–117. doi: 10.1109/TPDS.2012.98.

[13] RAVEENDRAN R and MENON R R. A novel aggregated statistical feature based accurate classification for internet traffic[C]. Proceedings of the 16 International Conference on Data Mining and Advanced Computing (SAPIENCE), Ernakulam, India, 2016: 225–232. doi: 10.1109/SAPIENCE.2016.7684123.

[14] MIAO Yuantian, RUAN Zichan, PAN Lei, *et al.* Comprehensive analysis of network traffic data[C]. 16th IEEE International Conference on Computer and Information Technology, Nadi, FIji, 2017: 423–430. doi: 10.1109/TPDS.2012.98.

[15] THAY C, VISOOTTIVISETH V, and MONGKOLLUKSAMEE S. P2P traffic classification for residential network[C]. Proceedings of the 2015 Computer Science and Engineering Conference (ICSEC), Chiang Mai, Thailand, 2015: 1–6. doi: 10.1109/ICSEC.2015.7401433.

[16] HUANG Yinxiang, LI Yun, and QIANG Baohua. Internet traffic classification based on min-max ensemble feature selection[C]. 2016 International Joint Conference on Neural Networks (IJCNN), Vancouver, Canada, 2016: 3485–3492. doi: 10.1109/IJCNN.2016.7727646.

[17] AUGUSTIN B and MELLOUK A. On traffic patterns of http applications[C]. Proceedings of the Global Telecommunications Conference (GLOBECOM 2011), Texas, USA, 2011: 1–6. doi: 10.1109/GLOCOM.2011.6134438.

[18] WANG Zaijian, DONG Yuning, *et al.* Internet video traffic classification using QoS features[C]. Proceedings of the 2016 International Conference on Computing, Networking and Communications (ICNC), Hawaii, USA, 2016: 1–5. doi: 10.1109/ICCNC.2016.7440599.

[19] SHAFIG M, YU X, and LAGHARI A A. WeChat text messages service flow traffic classification using machine learning technique[C]. Proceedings of the 6th International Conference on IT Convergence and Security (ICITCS), Prague, Czech, 2016: 1–5. doi: 10.1109/ICITCS.2016.7740379.

[20] DUBIN R, HADAR O, RICHMAN I, *et al.* Video quality representation classification of Safari encrypted DASH streams[C]. Proceedings of the 1st Digital Media Industry & Academic Forum (DMIAF). Santorini, Greece, 2016: 213–216. doi: 10.1109/DMIAF.2016.7574935.

[21] NOVAKOVIC J. Toward optimal feature selection using ranking methods and classification algorithms[J]. *Yugoslav Journal of Operations Research*, 2011, 21(1): 119–135. doi: 10.2298/YJOR1101119N.

[22] HALL M A. Correlation-based feature selection for machine learning[D]. [Ph.D. dissertation], The University of Waikato, 1999.

[23] KONONENKO I, ŠIMEC E, and ROBINK-ŠIKONJA M. Overcoming the myopia of inductive learning algorithms with RELIEFF[J]. *Applied Intelligence*, 1997, 7(1): 39–55. doi: 10.1023/A:1008280620621.

[24] Telecommunication Standardization Sector of ITU-2013, Parametric non-intrusive assessment of audiovisual media streaming quality[S]. 2013.

杨凌云： 女，1983 年生，博士生，讲师，研究方向为多媒体流分类.

董育宁： 男，1955 年生，博士生导师，教授，研究方向为多媒体通信.

王再见： 男，1980 年生，副教授，研究方向为多媒体流分类.