基于空时分形的网络流量认知模型

汤萍萍[1,2,3]，张晖[2]，董育宁[2]，董国青[1,2]
（1.安徽师范大学物理与电子信息学院，安徽 芜湖 241002；2.南京邮电大学通信与信息工程学院，江苏 南京 210003；
3.安徽华东光电技术研究所有限公司，安徽 芜湖 241002）

摘　要：为解决传统分形特征在网络流量认知精度和速度上难以兼具的问题，以分形理论为基础提出网络流量空时分离思想并创建空时分形特征，以此设计一种新型的流量认知模型——空时分形模型（SFM）。空时分形观测空间序列和时间序列，然后基于勒让德变换建立向量再折射到对偶空间形成特征。空时分形的物理含义在于从不同空间和时间尺度上获得数据突发特征，而传统分形是空时分形在空间和时间维度上的特征融合，空时分形相比传统分形刻画出更多细节特征，以此进行流量认知更为精准。此外，空时分形相比传统分形更易计算，使SFM在增强认知精度的同时提升认知速度。实验数据显示，SFM的认知性能优于其他方法。

关键词：网络流量；认知精度；认知速度；分形理论
中图分类号：TN919
文献标志码：A
DOI：10.11959/j.issn.1000–436x.2025091

Network traffic cognition model based on space-time fractals

TANG Pingping[1,2,3], ZHANG Hui[2], DONG Yuning[2], DONG Guoqing[1,2]
1. School of Physics and Electronic Information, Anhui Normal University, Wuhu 241002, China
2. College of Telecommunications and Information Engineering, Nanjing University of Posts and Telecommunications, Nanjing 210003, China
3. Anhui Huadong Photoelectric Technology Research Institute Co., Ltd., Wuhu 241002, China

Abstract: Considering the problem of traditional fractal (TF) features being difficult to achieve both high accuracy and fast speed in network traffic cognition, the idea of space-time separation was proposed on the basis of fractal theory. With space-time fractal (SF) features generated by the space-time separation, a new traffic cognition system called the space-time fractal model (SFM) was established. In order to obtain SF, the spatial and temporal sequences were observed, and further constructed to generate vectors by Legendre transformation, which were mapped into dual space. The physical significance of SF lied in capturing the characteristics of traffic bursts at different scales of space and time, while TF were the fusion of SF across spatial and temporal scales. Compared with TF, SF represented network traffic more comprehensively and thus were able to identify traffic more accurately. Moreover, SF were more computationally efficient than TF, enabling SFM to achieve high cognition speed as well as strong cognition accuracy. The experimental results show that the cognition performance of SFM is superior to other methods.

Keywords: network traffic, cognition accuracy, cognition speed, fractal theory

---

收稿日期：2025–03–10；**修回日期：**2025–05–09
基金项目：国家自然科学基金资助项目（No.62071005）；国家重点研发计划基金资助项目（No.2020YFB2104004）；江苏省重点研发计划基金资助项目（No.BE2021725）；安徽省自然科学基金资助项目（No.2308085Y02, No.2208085MF155）；安徽省高校自然科学基金资助项目（No.KJ2021A0124）；安徽省博士后科研基金资助项目（No.2024C946）
Foundation Items: The National Natural Science Foundation of China (No.62071005), The National Key Research and Development Program of China (No.2020YFB2104004), The Key Research and Development Program of Jiangsu Province (No.BE2021725), The Natural Science Foundation of Anhui Province (No.2308085Y02, No.2208085MF155), The Natural Science Foundation of the Higher Education Institutions of Anhui Province (No.KJ2021A0124), The Anhui Postdoctoral Scientific Research Program Foundation (No.2024C946)

0　引言

6G作为万物互联、万物智联的载体将通信、感知、计算、控制、信息和数据进行深度耦合[1-2]。在6G技术中，流量认知是进行区分服务、资源管理、网络监测、安全控制、态势感知等系列网络行为的重要基础[3-4]，业内权威组织及机构非常重视这一基础研究：国际标准化组织3GPP指出，流量认知是6G网络架构中执行区分服务的前提和基础[5]，如图1所示，流量认知部署在边缘路由器E_2上，在E_2中流量被分成文本、语音、视频等不同类别，然后进入相应的队列等待调度[6]。华为在其白皮书中阐述，流量认知是网络运行的基础，一切管理、调度、资源分配、传输处理等网络行为从流量认知开始[7]。

基于用户隐私和网络安全的需求，网络流量一般被加密传输。传统网络流量分类技术（如深度包检测[8]、应用层协议解析[9]、深度解码[10]）无法处理加密流量，于是基于网络流的流特征进行流量认知成为当前主流方法[11-12]。网络流是一组满足五元组<源IP，目的IP，源端口，目的端口，协议>的数据包的集合[13-17]，包含2类重要信息。

$$\{(P_i,T_i)|_{i=1,2,\cdots,N}\} \tag{1}$$

其中，P_i是第i个数据包的大小，T_i是该数据包与前一个数据包的间隔时间，N为数据包的个数。为直观感受网络流外貌，基于P_i和T_i计算单位时间I内的数据量[12]，可表示为

$$F_I=\int_{kI}^{(k+1)I} g(t)\,\mathrm{d}t \tag{2}$$

其中，数据传输速率函数$g(t)$由P_i和T_i相除再以分段函数呈现，如图2所示。图3为微信音频和电子邮件网络流，这里取I=10 ms，即每隔10 ms统计一次数据量，得到第1个I内的数据量（0~10 ms），第2个I内的数据量（10~20 ms），第3个I内的数据量（20~30 ms），…，第k个I内的数据量。此外，流量认知计算并不需要完整的网络流，只需截取开始的一段，为此将网络流分割成若干子流，第m个子流为

$$\{(P_i,T_i)|_{i=p+1,p+2,\cdots,p+N_m-1}\} \tag{3}$$

$$\text{s.t. } p=\sum_{i=1}^{m-1}N_i \tag{4}$$

其中，N_m表示第m个子流中包含的数据包个数。

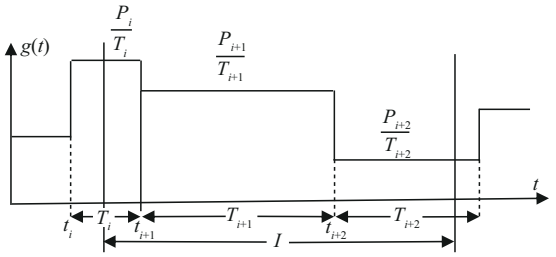


图2　F_I计算过程

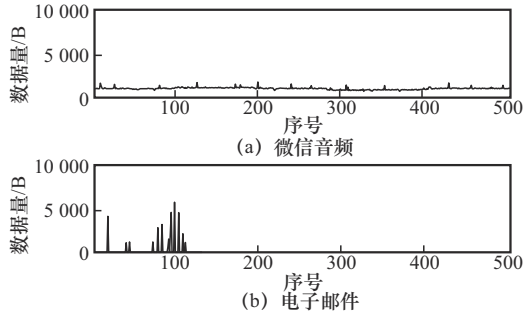


图3　网络流

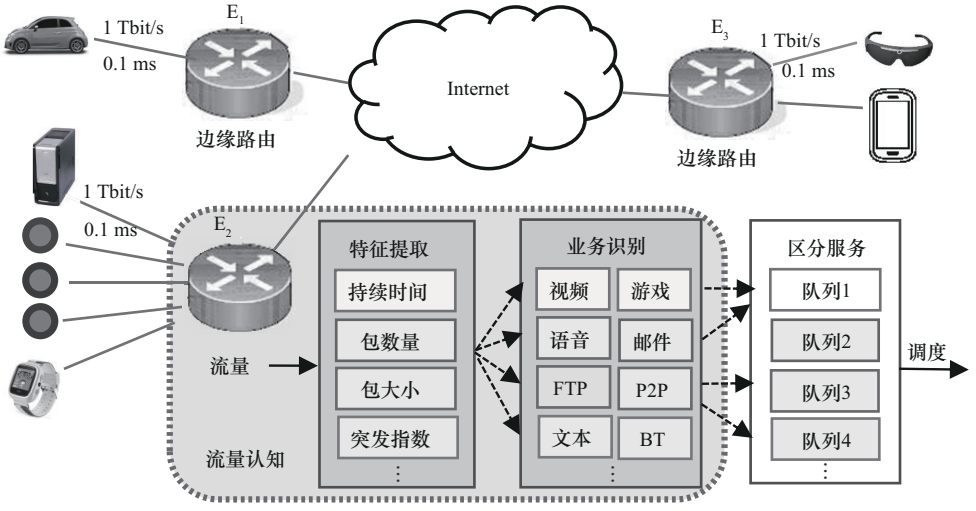


图1　流量认知模型（参考标准3GPP TS23.203）

众多研究表明[13]，应用程序发出的数据流总是遵循特定的协议、传输方式和对话规则，因此有迥异的统计特征，如生存期、平均包大小，这些统计特征可用来区分网络流类型。如图3所示，电子邮件业务流的生存期为1 s左右，微信音频一般至少数秒，文件传输协议（FTP）下载则用分来计时。包大小可以作为区分这3种流量的显著特征。但是，随着业务种类变细、流量种类变多，统计特征逐渐失效，不论是生存期还是包大小特征都难以对抖音视频和微信视频进行有效区分，更是无法区分优酷标清（SD）和高清（HD）等视频流量，如图4所示。为此，研究人员深入分析流量数据的传输协议和对话机制以发现固有的行为模式来判别流量类型，建立基于行为特征的流量认知方法[14]。例如，文献[15]根据同步序列编号（SYN）和确认字符（ACK）应答机制分析握手阶段数据包之间的关联特征并形成马尔可夫链来认知流量，对简单邮件传输协议（SMTP）、FTP、超文本传输协议（HTTP）等10种流量的平均识别准确率达到96%。文献[16]指出在连接握手阶段数据消息大小是固定的，网络流前几个包所包含的这种行为模式可作为关键特征来辨识Tor类型流量，准确率高达98%。基于行为特征的流量认知方法可靠稳定，但是行为特征针对性强，如识别端到端（P2P）流量的行为特征无法识别电子邮件，识别电子邮件的行为特征无法识别FTP流量。此外，行为特征是基于研究人员长期积累的先验知识，目前还没有使用行为特征对未知流、零日流（有研究表明零日流数量超过网络流量总数的50%[17]）进行分类识别的相关研究。

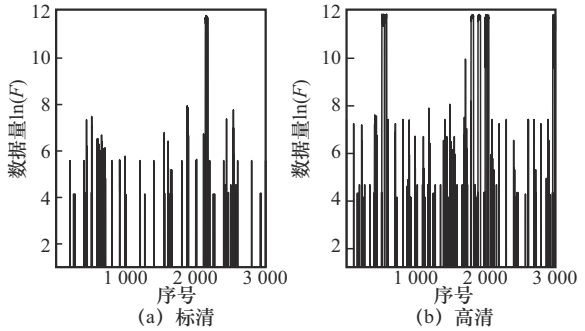


图4　优酷标清和高清视频流量

近年来，基于分形特征的流量认知方法成为本领域研究热点[18]。分形特征与行为特征类似，基于数据包之间的关联特性对流量类型进行认知，不同之处在于，分形特征从不同观测尺度分析并获取数据之间的相关性（数据突发特征），不需要深度挖掘网络应用背后的数据传输机制和控制策略模式，在异构网络环境中面对变化的目标类有较强的普适性，对新兴流、未知流、零日流也有较好的认知效果[19]。然而，传统分形特征的计算需要较高的时空复杂度，认知模型一般以牺牲速度的方式来获得认知精度，在精度和速度上难以兼得[20]。为此，本文在分形理论的基础上提出并论证空时分离思想，以此建立新型的流量认知模型——空时分形模型（SFM, space-time fractal model）以突破现有分形方法的局限。本文主要贡献概括如下。

1) 首次提出空时分离思想建立空时分形特征。不同于传统分形，空时分形是流量经空时分离后由勒让德变换折射到对偶空间而形成的分形特征，反映的是不同空间、时间尺度观察网络流得到的数据突发性，比传统分形刻画出更多细节特征，以此进行流量认知可提升模型的认知精度。

2) 提出一种相似度量方法：空时分形相似度。该相似度体现流量变化轨迹的相似程度，本文利用理论分析工具论证了这一相似度用于表述空时分形差异度的合理性，实验数据也进一步展示了空时分形相似度用于流量认知的适用性。

3) SFM突破已有流量认知方法的瓶颈和制约，实现高精准的快速流量认知。SFM仅需1 500个数据包即可获得足够细节特征对22种流量达到92%的认知准确率，而传统分形方法无法达到这种水平。SFM在提升认知精度的同时，促进认知计算速度大幅提升，实现认知精度和速度的同步优化。

1　相关工作

根据流特征的类型，当前的流量认知方法可划分为统计特征方法[12-13]、行为特征方法[15-16]以及分形特征方法[18-19]。

统计特征方法的大概过程是[21-23]：先对网络流数据进行大量观察，获得平均包大小、间隔时间、包数量、上下行字节数、流持续时间、平均速率等统计特征；再用特征选择算法去除冗余特征或提取关键特征；最后基于有效特征进行流量认知计算。文献[6]为区分流媒体和FTP流，提取平均包大小、速率等有效特征，然后映射为矢量距离，用K-means聚类找到类中心点。在线分类识别时，计算

待测流量与各类中心点的距离，选择距离最近的类。然而随着研究的深入，统计特征显露出以下局限性。1) 认知粒度有限。例如，文献[24]基于快速相关性依存过滤（FCBF）算法选择平均包大小、包峰值等 12 种特征，仅用于识别 P2P 流量。2) 适当增加特征会获得更高准确率，但计算和存储开销呈指数级增长（维数灾）[25]。此外，特征过多可能会导致过拟合，反而降低准确率[26-27]。3) 随着环境变化，特征也会变化，有效特征需要及时调整。文献[23]中指出，借助包采样方法可以加快分类识别效率，但是因有效信息减少导致特征变化使得识别准确率下降，于是有效特征需要调整，这又引起其他麻烦，例如，文献[14]基于决策树进行分类，增加或删除某个特征（对应决策树中的叶节点）可能导致整个树的更新，工作量巨大。研究者们将上述统计特征所面临的诸多问题统称为特征工程[28-29]。研究人员分析发现导致特征工程问题的主要原因在于其数据独立性假设，即，假设业务流中持续到达的包是相互独立的，数据包的大小是统计独立的，然而众多研究表明网络流量的传输一般遵循着特定的对话协议、应答机制、校验方式、重传策略、控制模式等，因此流量数据包之间具有较强的关联性，这种关联性可靠且稳定，可用于流量的分类识别，于是出现众多基于行为特征的流量认知方法。

基于行为特征的流量认知模型探索流量数据包之间的关联特征来认知流量，例如，文献[14]研究发现网络流由于传输策略和对话机制使得源 IP 和目的 IP 之间形成特定形状的辐射图，这些辐射图清晰地反映出流量数据包之间的对话行为特征，对于区分流媒体视频、现场直播视频、P2P 点播视频以及网络游戏达到 95% 的准确度。另外，处于传输控制协议（TCP）连接协商阶段的数据包顺序是预先定义的，于是文献[15]根据传输协议利用 SYN 和 ACK 的应答机制分析握手阶段数据包之间的关联特征并形成马尔可夫链来实现应用层协议的分类。P2P 对等节点在建立完传输层连接（如 TCP）之后便开始进行远端（Peer）的协议握手，文献[30]指出在 BT（BitTorrent）连接阶段数据对话模式具有固定的消息大小和方向，这些明显的行为模式可用来判断 P2P 的具体类型（如 Thunder 迅雷下载、PPTV 网络电视等），识别准确率均达到 97% 以上。文献[16]对数据包之间的关联性进行深度分析研究，发现不同类型的流量所发出的数据包大小之间呈现一定规律（有些文献称为时序特征[17]），这些数据包之间的顺序关联特征可有效用于流量认知。然而，行为特征来源于应用程序特定的数据传输机制和协议应答模式，这种特征虽然稳定可靠但是难于提炼，需要研究人员足够了解每种应用在运行中的协商机制，正因如此，1) 截至目前，基于行为特征的流量认知模型所设置的目标类数量有限，一般不超过 10 个；2) 行为特征针对特定应用提炼行为模式，针对性强、普适性弱，在泛在异构网络环境中面对变化的目标类适应能力差。随着 6G 万物互联技术发展，业务种类越来越多且持续变化，基于行为特征的流量认知面临巨大挑战。

近年来，一些研究者尝试使用分形特征来认知流量类型。分形特征是以数据相关性为核心、以随机数据的变化特性为内容，因此在非平稳的高可变网络中具有较好的认知效果，例如，Areström 等[18]使用分形指数 $D(h(q))$ 对视频流、社交网络、音频通信、网页浏览、文本通信以及批量下载 6 种流量进行分类识别，准确率达到 96%。文献[12]基于小波阈 Hurst 分形指数对 15 种流量进行认知，包括电驴（e Donkey）、Direct Connect、Gnutella、Fast Track 等，准确率高于统计特征方法和行为特征方法。文献[28]优化小波阈多重分形指数，使网络流在早期阶段就可以被有效识别，而不用等到流全部结束，极大促进了分形理论在流量认知领域的应用。文献[19]采用分形谱对网络流进行超细分类，将视频流量进一步分为标清和高清，如图 5 所示，这一研究充分展示了分形理论用于网络流量超细分类的可行性。然而，分形特征用于流量认知也存在局限。文献[28]虽然在早期就能对网络流量类型进行有效预判，但小波阈多重分形指数的计算过程复杂，不适合在线认知计算。文献[19]对 22 种流量实施精准认知，核阈 Q 越大分形细节特征越明显越有利于细分类，但分形谱的计算量随 Q 的增加直线上升。当前，分形方法基本难以同时提升认知精度和速度。为此，本文在传统分形的理论基础上，对流量进行空时分离并通过勒让德变换建立向量再折射到对偶空间形成空时分形特征。空时分形反映不同空间和时间尺度上的数据突发特性，而传统分形是空时分形在空间和时间维度上的特征融合。因此，

空时分形比传统分形描绘更多的细节特征，认知更加精准；此外，基于空时分离的空时分形相比于传统分形也更易计算，认知速度获得大幅提升。本文提出的基于空时分形的SFM实现认知精度和认知速度的同步优化。

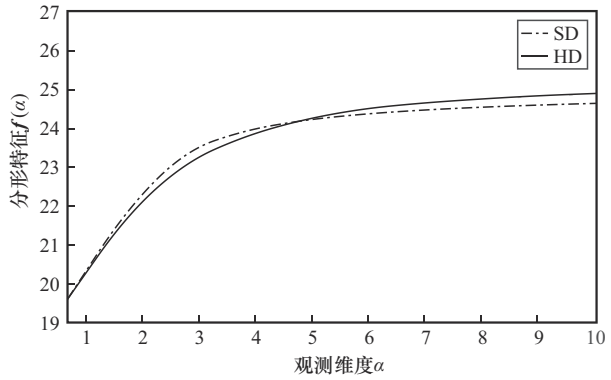


图5　优酷标清流量和高清流量的分形谱

2　空时分形模型

2.1　空时分离

在流量分形理论中，Leland等[31]证明了网络总体流量具有分形特性。总体流量分割成网络流F_I，网络流F_I汇聚成总体流量，Tang等[19]进一步论证网络流F_I具有分形特性。另外，众多研究已证明网络流的时间序列$\{T_i\}$呈现分形特性[32-33]。以下将证明网络流的空间序列$\{P_i\}$也呈现分形特性。

定理1　已知F_I、T具有分形特性，那么P也具有分形特性。

证明　P为T时间内传输的数据量，因此，有

$$\lim_{I\to T}\left(\frac{F_I}{I}\right)=\frac{P}{T} \tag{5}$$

也即

$$P=\lim_{I\to T}\left(\frac{F_IT}{I}\right) \tag{6}$$

由分形理论可知[31]，若事物X具备分形特性，那么观测尺度ε及其测度函数$N(\varepsilon)$满足

$$N(\varepsilon)\propto\varepsilon^{\boldsymbol{f}(\alpha)} \tag{7}$$

$$\text{s.t. } N(\varepsilon)=\frac{X}{X_\varepsilon} \tag{8}$$

其中，$\boldsymbol{f}(\alpha)$为事物X的分形特征，X_ε表示在观测尺度ε下观测事物X，α为观测维度。在文献[31]中，求解Hurst参数所用的测度函数$N(\varepsilon)$是基于观测对象的方差，也即，Hurst参数是多重分形特征$\boldsymbol{f}(\alpha)$取二阶矩的特定情况。由上，已知观测对象X具有分形特征$\boldsymbol{f}(\alpha)$，则测度函数$N(\varepsilon)$和$\boldsymbol{f}(\alpha)$满足式(7)。同理，已知F_I和T皆具有分形特性，那么$F_{\varepsilon I}$表示在观测尺度ε下观测F_I，T_ε表示在观测尺度ε下观测T，对应分形特征$\boldsymbol{f}_{F_I(\alpha)}$和$\boldsymbol{f}_T(\alpha)$，则有

$$N_{F_I}(\varepsilon)\propto\varepsilon^{\boldsymbol{f}_{F_I}(\alpha)} \tag{9}$$

$$N_T(\varepsilon)\propto\varepsilon^{\boldsymbol{f}_T(\alpha)} \tag{10}$$

由式(6)和式(8)可得

$$N_P(\varepsilon)=\frac{P}{P_\varepsilon}=\frac{\lim\limits_{I\to T}\left(\frac{F_IT}{I}\right)}{\lim\limits_{I\to T}\left(\frac{F_{\varepsilon I}T_\varepsilon}{I}\right)}=\lim_{I\to T}\left(\frac{F_I}{F_{\varepsilon I}}\cdot\frac{T}{T_\varepsilon}\right)=$$
$$\lim_{I\to T}N_{F_I}(\varepsilon)N_T(\varepsilon) \tag{11}$$

根据分形理论，由式(9)和式(10)进一步可得

$$\boldsymbol{f}_{F_I}(\alpha)=\lim_{\varepsilon\to0}\frac{\ln N_{F_I}(\varepsilon)}{\ln\varepsilon} \tag{12}$$

$$\boldsymbol{f}_T(\alpha)=\lim_{\varepsilon\to0}\frac{\ln N_T(\varepsilon)}{\ln\varepsilon} \tag{13}$$

基于式(12)和式(13)，定义$\boldsymbol{f}_P(\alpha)$函数为

$$\boldsymbol{f}_P(\alpha)=\lim_{\varepsilon\to0}\frac{\ln N_P(\varepsilon)}{\ln\varepsilon} \tag{14}$$

接下来，求解未知的$\boldsymbol{f}_P(\alpha)$。由式(11)可得

$$\boldsymbol{f}_P(\alpha)=\lim_{\varepsilon\to0}\frac{\ln N_P(\varepsilon)}{\ln\varepsilon}=$$
$$\lim_{\varepsilon\to0}\frac{\ln\lim\limits_{I\to T}(N_{F_I}(\varepsilon)N_T(\varepsilon))}{\ln\varepsilon}=$$
$$\lim_{I\to T}\left(\lim_{\varepsilon\to0}\frac{\ln(N_{F_I}(\varepsilon)N_T(\varepsilon))}{\ln\varepsilon}\right)=$$
$$\lim_{I\to T}\left(\lim_{\varepsilon\to0}\frac{\ln N_{F_I}(\varepsilon)+\ln N_T(\varepsilon)}{\ln\varepsilon}\right)=$$
$$\lim_{I\to T}\left(\lim_{\varepsilon\to0}\frac{\ln N_{F_I}(\varepsilon)}{\ln\varepsilon}+\lim_{\varepsilon\to0}\frac{\ln N_T(\varepsilon)}{\ln\varepsilon}\right)=$$
$$\lim_{I\to T}(\boldsymbol{f}_{F_I}(\alpha)+\boldsymbol{f}_T(\alpha))=\lim_{I\to T}\boldsymbol{f}_{F_I}(\alpha)+\boldsymbol{f}_T(\alpha)$$

如果F_I、T具有分形特性，即存在$\boldsymbol{f}_{F_I}(\alpha)$和$\boldsymbol{f}_T(\alpha)$，因此$I\to T$时$\boldsymbol{f}_{F_I}(\alpha)$存在。由此，空间序列$P$存在分形特性，其分形特征取决于$\boldsymbol{f}_{F_I}(\alpha)$和$\boldsymbol{f}_T(\alpha)$。

如图6所示，传统分形方法是基于F_I计算分形特征$\boldsymbol{f}_{F_I}(\alpha)$对网络流进行分类识别，先基于$P$、$T$由式(2)得到$F_I$，再计算得到$\boldsymbol{f}_{F_I}(\alpha)$。由图2所示$F_I$计算过程可知，$F_I$的计算比较耗时，基于$F_I$的传统

分形特征的计算量较大。既然流量在P和T这2个维度上分别具有分形特性，因此本文直接基于P、T维序列分别计算其相应的分形特征形成空时分形对网络流实施认知计算，大幅减少计算量。

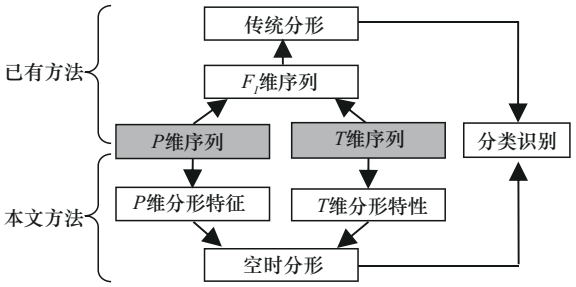


图6　传统分形与空时分形

P维序列、T维序列皆具有分形特性，那么理论上可依据式(7)分别计算其对应的分形特征$\boldsymbol{f}_P(\alpha)$和$\boldsymbol{f}_T(\alpha)$，但测度函数$N(\varepsilon)$的计算过程较为烦琐[31]，在实践中$\boldsymbol{f}(\alpha)$常用数值估计的方法来获得近似值，如勒让德变换下的无偏估计[33]：设$X=\{X_i, i=1,2,\cdots,N\}$为离散随机序列且具备分形特性。将离散序列$\{X(i)\}$划分成$m$个不相重叠的块，对这些块进行合并操作得到$m$阶聚并序列。

$$X^{(m)}=\left\{\sum_{i=mn-m+1}^{mn} X_i,\ n=0,1,\cdots,\frac{N}{m}\right\} \tag{15}$$

对m阶聚并序列进行q阶计算并求和

$$S_m(q)\triangleq\sum_{k=1}^{\frac{N}{m}}\left|X^{(m)}(k)\right|^q \tag{16}$$

最后得到分形特征序列

$$\boldsymbol{f}(\alpha)\triangleq\tau(q)=\lim_{m\to\infty}\frac{\ln S_m(q)}{\ln m} \tag{17}$$

其中，$\tau(q)$为勒让德变换下的分形估计谱，是离散序列，用$\boldsymbol{f}(\alpha)$表示离散的分形特征，本文使用$\tau(q)$来快速计算网络流在空间尺度和时间尺度上的分形特征。

2.2　空时分形

将空间序列$\{P_i\}$和时间序列$\{T_i\}$分别代入式(15)~式(17)中，从而形成空间和时间2个维度上的分形特征序列$\boldsymbol{f}_P(\alpha)$和$\boldsymbol{f}_T(\alpha)$。前者描述流量包大小的变化特征；后者描述流量包在时间上的突发特征。如式(18)所示，将特征序列折射到对偶空间逐行点乘，其数值对应的物理含义是，不同空间和时间尺度上网络流数据突发量的变化特征。

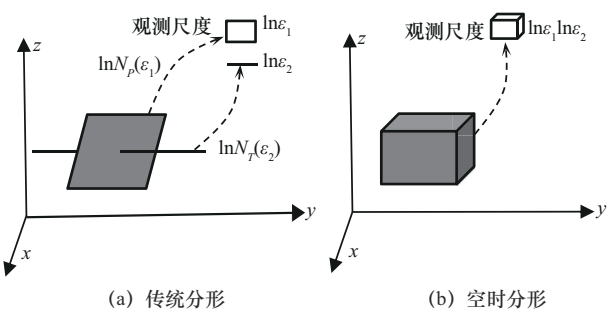
$$\boldsymbol{M}=\boldsymbol{f}_P(\alpha)*\boldsymbol{f}_T(\alpha)^{\mathrm{T}} \tag{18}$$

其中，$\boldsymbol{f}_P(\alpha)$是基于空间序列P建立的分形特征，观测尺度最小为q=1，最大为$q=\lceil \mathrm{lb}\,N\rceil$。同理$\boldsymbol{f}_T(\alpha)$是时间序列$T$对应的分形特征。$P$和$T$这2个维度上的分形特征描述的是当观测尺度$q$从1到$\lceil \mathrm{lb}\,N\rceil$变化时，流量数据在时间和空间上所呈现出的变化轨迹。

综上所述，传统分形的观测对象是流量F_I，揭示流量数据突发特征，对应$\boldsymbol{f}_{F_I}(\alpha)$。空时分形的观测对象是包大小$P_i$和间隔时间$T_i$，分别揭示包大小和包间隔的突发特征，对应$\boldsymbol{f}_P(\alpha)$和$\boldsymbol{f}_T(\alpha)$。根据式(13)和式(14)可得

$$M=\lim_{\varepsilon\to 0}\frac{\ln N_P(\varepsilon)}{\ln\varepsilon}\lim_{\varepsilon\to 0}\frac{\ln N_T(\varepsilon)}{\ln\varepsilon}=\lim_{\varepsilon_1\to 0}\lim_{\varepsilon_2\to 0}\frac{\ln N_P(\varepsilon_1)\ln N_T(\varepsilon_2)}{\ln\varepsilon_1\ln\varepsilon_2}$$

如图7所示，为便于理解，包大小的突发量$\ln N_P(\varepsilon_1)$用一个菱形方块来表示，其对应的观测尺度$\ln\varepsilon_1$也是一个小方块，包间隔的突发量$\ln N_T(\varepsilon_2)$用一条线段来表示，其观测尺度$\ln\varepsilon_2$也是一条线段；$\ln N_P(\varepsilon_1)$乘以$\ln N_T(\varepsilon_2)$是一个长方体，其观测尺度$\ln\varepsilon_1\ln\varepsilon_2$也是一个小长方体。即空时分形揭示的是体量的突发特征，传统分形揭示的也是体量的突发特征（单位时间I内的数据量，流量即体量），不同之处在于，传统分形的观测对象是融合后的$\ln N_{F_I}(\varepsilon)$，观测尺度也是融合后的$\ln\varepsilon$；空时分形的2个观测对象是分开的。类似于$d=\sqrt{x^2+y^2}$，d相当于传统分形，x和y相当于空时分形，x和y融合成d会丢失细节特征。本文提出的空时分形反映不同空间尺度和时间尺度上的数据突发特征，传统分形是空时分形在空间尺度和时间尺度上的特征融合，空时分形相比传统分形刻画出更多细节特征。



(a) 传统分形　　(b) 空时分形

图7　空时分形物理意义

定理2 空时分形$\boldsymbol{M}$唯一标识网络流。

证明 为此先固定时间序列的观测尺度$\boldsymbol{f}_T(\alpha)|_{\alpha=q'}$，仅观测空间变化特征$\boldsymbol{f}_P(\alpha)$。对于流$F_I^a$，其对应的空间变化特征为$\boldsymbol{f}_{Pa}(\alpha)$，流$F_I^b$对应为$\boldsymbol{f}_{Pb}(\alpha)$，求解聚集流$F_I{}^c=F_I{}^a+F_I{}^b$的空间变化特征$\boldsymbol{f}_{Pc}(\alpha)$如下。根据流量分形理论，由式(8)可得

$$N_{Pc}(\varepsilon)=\frac{P_c}{P_{c\varepsilon}}=\frac{P_a+P_b}{P_{a\varepsilon}+P_{b\varepsilon}}=\frac{P_a}{P_{a\varepsilon}}\cdot\frac{1}{1+\dfrac{P_{b\varepsilon}}{P_{a\varepsilon}}}+\frac{P_b}{P_{b\varepsilon}}\cdot\frac{1}{1+\dfrac{P_{a\varepsilon}}{P_{b\varepsilon}}}$$

其中，$P_{a\varepsilon}$和$P_{b\varepsilon}$大于0，且$\lim\limits_{\varepsilon\to0}P_{a\varepsilon}$和$\lim\limits_{\varepsilon\to0}P_{b\varepsilon}$为同级无穷小，因此可得

$$\lim_{\varepsilon\to0}N_{Pc}(\varepsilon)=\lim_{\varepsilon\to0}\left(k_1N_{Pa}(\varepsilon)+k_2N_{Pb}(\varepsilon)\right)\tag{19}$$

求极限后可知$k_1+k_2=1$，$\sqrt{k_1k_2}<\frac{1}{2}$。再由式(7)可得

$$N_{Pa}(\varepsilon)\propto\varepsilon^{f_{Pa}(\alpha)}\tag{20}$$

$$N_{Pb}(\varepsilon)\propto\varepsilon^{f_{Pb}(\alpha)}\tag{21}$$

由式(14)和式(19)可得

$$\boldsymbol{f}_{Pc}(\alpha)=\lim_{\varepsilon\to0}\frac{\ln\left(k_1N_{Pa}(\varepsilon)+k_2N_{Pb}(\varepsilon)\right)}{\ln\varepsilon}\tag{22}$$

其中，$N_{Pa}(\varepsilon)$和$N_{Pb}(\varepsilon)$是以尺度ε观测2条网络流获得的观测值，都是大于0的正数，因此，有

$$k_1N_{Pa}(\varepsilon)+k_2N_{Pb}(\varepsilon)>2\sqrt{k_1k_2N_{Pa}(\varepsilon)N_{Pb}(\varepsilon)}\tag{23}$$

$$2\max(N_{Pa}(\varepsilon),N_{Pb}(\varepsilon))>k_1N_{Pa}(\varepsilon)+k_2N_{Pb}(\varepsilon)\tag{24}$$

因为$\lim\limits_{\varepsilon\to0}\ln\varepsilon$是负数，那么由式(22)和式(23)可得

$$\boldsymbol{f}_{Pc}(\alpha)<\lim_{\varepsilon\to0}\frac{\ln2\sqrt{k_1k_2N_{Pa}(\varepsilon)N_{Pb}(\varepsilon)}}{\ln\varepsilon}<\frac{1}{2}\lim_{\varepsilon\to0}\frac{\ln N_{Pa}(\varepsilon)N_{Pb}(\varepsilon)}{\ln\varepsilon}=\frac{1}{2}\left(\lim_{\varepsilon\to0}\frac{\ln N_{Pa}(\varepsilon)}{\ln\varepsilon}+\lim_{\varepsilon\to0}\frac{\ln N_{Pb}(\varepsilon)}{\ln\varepsilon}\right)=\frac{1}{2}\left(\boldsymbol{f}_{Pa}(\alpha)+\boldsymbol{f}_{Pb}(\alpha)\right)$$

被记为

$$\sup\left(\boldsymbol{f}_{Pc}(\alpha)\right)=\frac{1}{2}\left(\boldsymbol{f}_{Pa}(\alpha)+\boldsymbol{f}_{Pb}(\alpha)\right)$$

由式(22)和式(24)可得

$$\boldsymbol{f}_{Pc}(\alpha)>\lim_{\varepsilon\to0}\frac{\ln2\max(N_{Pa}(\varepsilon),N_{Pb}(\varepsilon))}{\ln\varepsilon}$$

若$N_{Pa}(\varepsilon)>N_{Pb}(\varepsilon)$，则$\boldsymbol{f}_{Pa}(\alpha)<\boldsymbol{f}_{Pb}(\alpha)$，且有

$$\lim_{\varepsilon\to0}\frac{\ln2\max(N_{Pa}(\varepsilon),N_{Pb}(\varepsilon))}{\ln\varepsilon}=\lim_{\varepsilon\to0}\frac{\ln2N_{Pa}(\varepsilon)}{\ln\varepsilon}=\boldsymbol{f}_{Pa}(\alpha)$$

若$N_{Pa}(\varepsilon)<N_{Pb}(\varepsilon)$，则$\boldsymbol{f}_{Pa}(\alpha)>\boldsymbol{f}_{Pb}(\alpha)$，且有

$$\lim_{\varepsilon\to0}\frac{\ln2\max(N_{Pa}(\varepsilon),N_{Pb}(\varepsilon))}{\ln\varepsilon}=\lim_{\varepsilon\to0}\frac{\ln2N_{Pb}(\varepsilon)}{\ln\varepsilon}=\boldsymbol{f}_{Pb}(\alpha)$$

综上可得

$$\inf\left(\boldsymbol{f}_{Pc}(\alpha)\right)=\min\left(\boldsymbol{f}_{Pa}(\alpha),\boldsymbol{f}_{Pb}(\alpha)\right)$$

在真实场景中，分形特征$\boldsymbol{f}(\alpha)$以正数形式存在，需要做处理：$-\boldsymbol{f}(\alpha)\to\boldsymbol{f}(\alpha)$。负数变正数后，上界变下界、最小值变最大值，于是得到

$$\inf\left(\boldsymbol{f}_{Pc}(\alpha)\right)=\frac{1}{2}\left(\boldsymbol{f}_{Pa}(\alpha)+\boldsymbol{f}_{Pb}(\alpha)\right)$$

$$\sup\left(\boldsymbol{f}_{Pc}(\alpha)\right)=\max\left(\boldsymbol{f}_{Pa}(\alpha),\boldsymbol{f}_{Pb}(\alpha)\right)$$

特别地，当$\boldsymbol{f}_{Pa}(\alpha)=\boldsymbol{f}_{Pb}(\alpha)=\boldsymbol{f}_P(\alpha)$时，则$\inf\left(\boldsymbol{f}_{Pc}(\alpha)\right)=\sup\left(\boldsymbol{f}_{Pc}(\alpha)\right)=\boldsymbol{f}_P(\alpha)$，物理意义是：若流$F_i^a$与流$F_i^b$为同类型，则聚集流仍为此类型；若流$F_i^a$与$F_i^b$为不同类型，那么聚集后空间变化特征既不是$F_i^a$也不是$F_i^b$。由此论证，在时间序列的观测尺度$\boldsymbol{f}_T(\alpha)|_{\alpha=q'}$下，特征$\boldsymbol{f}_{Pa}(\alpha)$具有唯一性，因此全部特征序列分量的正交矩阵$\boldsymbol{M}=\boldsymbol{f}_P(\alpha)*\boldsymbol{f}_T(\alpha)^{\mathrm{T}}$可唯一标记网络流。

2.3 空时分形相似度

空时分形$\boldsymbol{M}$描述的是随着观测尺度的变化流量突发特征的变化轨迹。一种类型的流量总是遵循特定的协议、传输方式，因此有着相似的变化轨迹，反映流量的固有特性。本文基于空时分形$\boldsymbol{M}$考量网络流相似度，实现对网络流的认知。为此，基于矩阵关系为空时分形$\boldsymbol{M}$定义相似度量

$$E(\boldsymbol{M}_a,\boldsymbol{M}_b)=\frac{\boldsymbol{M}_a\boldsymbol{M}_b{}^{\mathrm{T}}+\boldsymbol{M}_b\boldsymbol{M}_a{}^{\mathrm{T}}}{\boldsymbol{M}_a\boldsymbol{M}_a{}^{\mathrm{T}}+\boldsymbol{M}_b\boldsymbol{M}_b{}^{\mathrm{T}}}\tag{25}$$

其中，$\boldsymbol{M}_a$和$\boldsymbol{M}_b$分别表示流F_i^a和F_i^b的空时分形。

定理3 $E(\boldsymbol{M}_a,\boldsymbol{M}_b)$可以准确提取$\boldsymbol{M}_a$和$\boldsymbol{M}_b$之间的相似性。

证明 为证明这一点，首先构造辅助函数

$$M(\boldsymbol{\beta})=\frac{\boldsymbol{\beta}^{\mathrm{T}}\left(\boldsymbol{M}_a\boldsymbol{M}_b{}^{\mathrm{T}}+\boldsymbol{M}_b\boldsymbol{M}_a{}^{\mathrm{T}}\right)\boldsymbol{\beta}}{\boldsymbol{\beta}^{\mathrm{T}}\left(\boldsymbol{M}_a\boldsymbol{M}_a{}^{\mathrm{T}}+\boldsymbol{M}_b\boldsymbol{M}_b{}^{\mathrm{T}}\right)\boldsymbol{\beta}}\tag{26}$$

设其共轭向量$\bar{\boldsymbol{\beta}}$使式(26)取得最大值，即

$$M(\bar{\boldsymbol{\beta}})=\max_{\boldsymbol{\beta}}\frac{\boldsymbol{\beta}^{\mathrm{T}}\left(\boldsymbol{M}_a\boldsymbol{M}_b{}^{\mathrm{T}}+\boldsymbol{M}_b\boldsymbol{M}_a{}^{\mathrm{T}}\right)\boldsymbol{\beta}}{\boldsymbol{\beta}^{\mathrm{T}}\left(\boldsymbol{M}_a\boldsymbol{M}_a{}^{\mathrm{T}}+\boldsymbol{M}_b\boldsymbol{M}_b{}^{\mathrm{T}}\right)\boldsymbol{\beta}}\tag{27}$$

由式(27)可知，共轭矩阵满足$\bar{\boldsymbol{M}}(\bar{\boldsymbol{\beta}})=\boldsymbol{M}(\bar{\boldsymbol{\beta}})$，因此有

$$1-\boldsymbol{M}(\bar{\boldsymbol{\beta}})=1-\max_{\boldsymbol{\beta}}\frac{\boldsymbol{\beta}^{\mathrm{T}}\left(\boldsymbol{M}_a\boldsymbol{M}_b^{\mathrm{T}}+\boldsymbol{M}_b\boldsymbol{M}_a^{\mathrm{T}}\right)\boldsymbol{\beta}}{\boldsymbol{\beta}^{\mathrm{T}}\left(\boldsymbol{M}_a\boldsymbol{M}_a^{\mathrm{T}}+\boldsymbol{M}_b\boldsymbol{M}_b^{\mathrm{T}}\right)\boldsymbol{\beta}}=\min_{\boldsymbol{\beta}}\frac{\boldsymbol{\beta}^{\mathrm{T}}(\boldsymbol{M}_a-\boldsymbol{M}_b)(\boldsymbol{M}_a-\boldsymbol{M}_b)^{\mathrm{T}}\boldsymbol{\beta}}{\boldsymbol{\beta}^{\mathrm{T}}(\boldsymbol{M}_a\boldsymbol{M}_a^{\mathrm{T}}+\boldsymbol{M}_b\boldsymbol{M}_b^{\mathrm{T}})\boldsymbol{\beta}} \tag{28}$$

也就是

$$\frac{\bar{\boldsymbol{\beta}}^{\mathrm{T}}(\boldsymbol{M}_a-\boldsymbol{M}_b)(\boldsymbol{M}_a-\boldsymbol{M}_b)^{\mathrm{T}}\bar{\boldsymbol{\beta}}}{\bar{\boldsymbol{\beta}}^{\mathrm{T}}(\boldsymbol{M}_a\boldsymbol{M}_a^{\mathrm{T}}+\boldsymbol{M}_b\boldsymbol{M}_b^{\mathrm{T}})\bar{\boldsymbol{\beta}}}=\min_{\boldsymbol{\beta}}\frac{\boldsymbol{\beta}^{\mathrm{T}}(\boldsymbol{M}_a-\boldsymbol{M}_b)(\boldsymbol{M}_a-\boldsymbol{M}_b)^{\mathrm{T}}\boldsymbol{\beta}}{\boldsymbol{\beta}^{\mathrm{T}}\left(\boldsymbol{M}_a\boldsymbol{M}_a^{\mathrm{T}}+\boldsymbol{M}_b\boldsymbol{M}_b^{\mathrm{T}}\right)\boldsymbol{\beta}} \tag{29}$$

对于任意向量$\boldsymbol{Z}$，有$\|\boldsymbol{Z}\|=\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{Z}$，因此式(29)的左边可以变换为

$$\frac{\bar{\boldsymbol{\beta}}^{\mathrm{T}}(\boldsymbol{M}_a-\boldsymbol{M}_b)(\boldsymbol{M}_a-\boldsymbol{M}_b)^{\mathrm{T}}\bar{\boldsymbol{\beta}}}{\bar{\boldsymbol{\beta}}^{\mathrm{T}}\left(\boldsymbol{M}_a\boldsymbol{M}_a^{\mathrm{T}}+\boldsymbol{M}_b\boldsymbol{M}_b^{\mathrm{T}}\right)\bar{\boldsymbol{\beta}}}=\frac{\bar{\boldsymbol{\beta}}^{\mathrm{T}}(\boldsymbol{M}_a-\boldsymbol{M}_b)(\boldsymbol{M}_a-\boldsymbol{M}_b)^{\mathrm{T}}\bar{\boldsymbol{\beta}}}{\bar{\boldsymbol{\beta}}^{\mathrm{T}}\boldsymbol{M}_a\boldsymbol{M}_a^{\mathrm{T}}\bar{\boldsymbol{\beta}}+\bar{\boldsymbol{\beta}}^{\mathrm{T}}\boldsymbol{M}_b\boldsymbol{M}_b^{\mathrm{T}}\bar{\boldsymbol{\beta}}}=\frac{\left\|(\boldsymbol{M}_a-\boldsymbol{M}_b)^{\mathrm{T}}\bar{\boldsymbol{\beta}}\right\|}{\left\|\boldsymbol{M}_a^{\mathrm{T}}\bar{\boldsymbol{\beta}}\right\|+\left\|\boldsymbol{M}_b^{\mathrm{T}}\bar{\boldsymbol{\beta}}\right\|}$$

对式(29)右边做同样操作，有

$$\frac{\left\|(\boldsymbol{M}_a-\boldsymbol{M}_b)^{\mathrm{T}}\bar{\boldsymbol{\beta}}\right\|}{\left\|\boldsymbol{M}_a{}^{\mathrm{T}}\bar{\boldsymbol{\beta}}\right\|+\left\|\boldsymbol{M}_b{}^{\mathrm{T}}\bar{\boldsymbol{\beta}}\right\|}=\min_{\boldsymbol{\beta}}\frac{\left\|(\boldsymbol{M}_a-\boldsymbol{M}_b)^{\mathrm{T}}\bar{\boldsymbol{\beta}}\right\|}{\left\|\boldsymbol{M}_a{}^{\mathrm{T}}\bar{\boldsymbol{\beta}}\right\|+\left\|\boldsymbol{M}_b{}^{\mathrm{T}}\boldsymbol{\beta}\right\|} \tag{30}$$

其中，$\bar{\boldsymbol{\beta}}$为一个向量，那么$\boldsymbol{M}_a{}^{\mathrm{T}}\bar{\boldsymbol{\beta}}$、$\boldsymbol{M}_b{}^{\mathrm{T}}\bar{\boldsymbol{\beta}}$就是矩阵$\boldsymbol{M}_a$和$\boldsymbol{M}_b$在$\bar{\boldsymbol{\beta}}$上的投影，$(\boldsymbol{M}_a-\boldsymbol{M}_b)^{\mathrm{T}}\bar{\boldsymbol{\beta}}$就是矩阵$\boldsymbol{M}_a$和$\boldsymbol{M}_b$的差异在$\bar{\boldsymbol{\beta}}$上的投影。因此式(28)意味着，基于式(25)的相似度量存在一个向量$\bar{\boldsymbol{\beta}}$，使$\boldsymbol{M}_a$和$\boldsymbol{M}_b$在该向量上的投影差异最小，由此证明$E(\boldsymbol{M}_a,\boldsymbol{M}_b)$可准确提取$\boldsymbol{M}_a$和$\boldsymbol{M}_b$之间的相似性。对于相似矩阵$\boldsymbol{A}$和$\boldsymbol{P}^{-1}\boldsymbol{A}\boldsymbol{P}$，可知$\mathrm{tr}(\boldsymbol{P}^{-1}\boldsymbol{A}\boldsymbol{P})=\mathrm{tr}(\boldsymbol{P}\boldsymbol{P}^{-1}\boldsymbol{A})=\mathrm{tr}(\boldsymbol{A})$，这里，tr(•)为矩阵的迹，相似矩阵具有相同的迹。由式(18)可知，$\boldsymbol{M}$是空间序列的分形特征$\boldsymbol{f}_P(\alpha)$与时间序列的分形特征$\boldsymbol{f}_T(\alpha)$的乘积，因此$\mathrm{tr}(\boldsymbol{f}_P(\alpha)\boldsymbol{f}_T(\alpha)^{\mathrm{T}})=\boldsymbol{f}_T(\alpha)^{\mathrm{T}}\boldsymbol{f}_P(\alpha)$，将式(25)相似度量的矢量矩阵转换成标量，并称之为空时分形相似度。

$$\mathrm{Sim}(\boldsymbol{M}_a,\boldsymbol{M}_b)\triangleq\frac{\mathrm{tr}(\boldsymbol{M}_a\boldsymbol{M}_b^{\mathrm{T}}+\boldsymbol{M}_b\boldsymbol{M}_a^{\mathrm{T}})}{\mathrm{tr}(\boldsymbol{M}_a\boldsymbol{M}_a^{\mathrm{T}}+\boldsymbol{M}_b\boldsymbol{M}_b^{\mathrm{T}})} \tag{31}$$

由式(31)可得，$\mathrm{Sim}(\boldsymbol{M}_a,\boldsymbol{M}_b)=\mathrm{Sim}(\boldsymbol{M}_b,\boldsymbol{M}_a)$，且Sim(·)范围在0到1之间，值越大说明两者相似度越高，极端情况下$\mathrm{Sim}(\boldsymbol{M}_a,\boldsymbol{M}_a)=1$，两者完全一致。

2.4　分类

分类过程借鉴文献[19]中基于K-means的分类器设计方法。设当前有L个类$\{\boldsymbol{M}_l\}_{l=1}^{L}$，每个类有若干条流$\{\cdots,F_I^j,F_I^k,\cdots\}$，$F_I^j$和$F_I^k$分别代表第$j$条流和第$k$条流。中心点记为$\{P_l\}_{l=1}^{L}$。Sim(·)服从0~1上的均匀分布，因此中心点由式(32)确定。

$$P_l\triangleq\min_{F_k\in\mathcal{M}_l}\left\{\max_{j\neq k,F_j\in\mathcal{M}_l}\mathrm{Sim}(\boldsymbol{M}_j,\boldsymbol{M}_k)\right\} \tag{32}$$

其中，$\boldsymbol{M}_j$和$\boldsymbol{M}_k$分别代表第j条流和第k条流对应的空时分形特征。由此可知，中心点P_l与类内其他点$\{\cdots,F_I^j,F_I^k,\cdots\}$的相似度均为一个比较小的量。对网络流$F_I^a$进行分类时，计算该条流与各中心点的相似度$\mathrm{Sim}(\boldsymbol{M}_a,\boldsymbol{M}_{P_l})$，选择最相似的进行以下操作

$$\mathrm{Be}(F_I^a,P_l)\triangleq\begin{cases}\in,\mathrm{Sim}(\boldsymbol{M}_a,\boldsymbol{M}_{P_l})\geqslant T\\ \notin,\mathrm{Sim}(\boldsymbol{M}_a,\boldsymbol{M}_{P_l})<T\end{cases} \tag{33}$$

其中，网络流F_I^a与中心点P_l的相似度若大于或等于阈值T，那么F_I^a属于类P_l；若相似度小于阈值T，那么F_I^a不属于类P_l。

3　实验

实验软件环境：用Wireshark捕捉实时业务流，用MATLAB R2016a仿真工具验证SFM的有效性。硬件环境为Win10 professional (64 bit/SP1), Intel (R) Core (TM) i7-7500U @ 2.70 GHz, 8 GB内存。实验使用的数据集如下。1) 在南京邮电大学校园网内获取的NJUPT数据集，该数据集包括6个类别：流媒体视频、即时音频（VoIP）、网页浏览、FTP传输、电子邮件、网络游戏。2) 因特网流量数据集UNB ISCX Network Traffic[34]，包含众多应用程序的流量数据，如Vimeo、YouTube、ICQ、Skype，脸书（Facebook）、Bittorrent等。该数据集的流量被分成8个类别。3) 在中国移动某地区数据中心采集的ISP数据集，整合了22种类型的流量，如视频点播、多人游戏、远程监控、线上诊疗等。

3.1　重要参数设置

SFM有N和q这2个重要参数。首先，关于解析度N的设置。分形特征的计算实际是在提取数据突发特性，为了获取足够稳定的特征，解析度N的设置不能过小。但是太大也没必要，N设置过大除了增加计算量，并不会增加新的显著突发特征。针对本文数据集，N设定为1 500，只需要1 500个数据包就可以获得足够稳定的特征。这里以视频流为

例分析N对空时分形特征的影响。对流序列分别取N_i={10 000, 7 000, 4 000, 2 000, 1 500, 1 200, 1 000, 800}，计算这些子流对应的空时分形，然后统计其对应的相似度Sim($\boldsymbol{M}_j$,$\boldsymbol{M}_k$)，结果如图8所示。

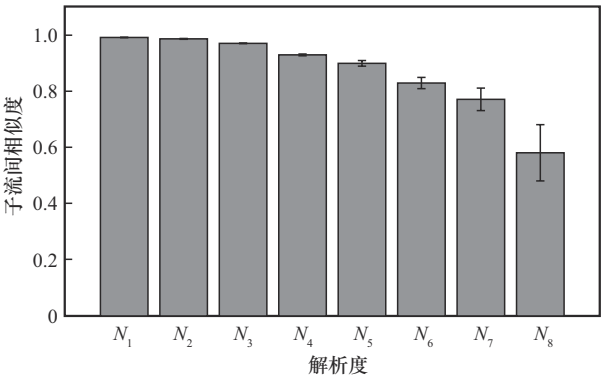


图8　子流间相似度

当N=N_1=10 000时，子流间矩阵相似度Sim($\boldsymbol{M}_j$, $\boldsymbol{M}_k$)~0.984±0.006，非常稳定。降低N_i稳定性变差，当N=N_8=800，Sim($\boldsymbol{M}_j$,$\boldsymbol{M}_k$)~0.598±0.127，子流间差异相当大，已经不能用作认知计算。对于其他类型流量做反复实验，情况大抵相似，因此本文最终选择解析度N=N_5=1 500，既保证分类识别的稳定性，计算量和存储量又不至于过大。当面对其他不同数据集（如攻击流），N可依据上述方法进行设置。

此外，关于观测尺度q。观测尺度最小为q=1，最大为q=⌈lb$N$⌉。由上所述，解析度N的物理意义在于“基于足量数据获得稳定的数据突发特征”，而观测尺度q的物理意义在于“基于足量的角度去观测流量以获得丰富的数据突发特征”，当q从1到⌈lb$N$⌉变化时，是在变换不同的观测角度去分析流量数据，其中，最为典型的是取q=2（二阶矩方差），对应的$f(\alpha)$即为Hurst指数。此外，较小的q（相当于低维视角）揭示的是事物的宏观突发特征，较大的q（相当于高维视角）揭示的是事物的微观突发特征。在粗粒度认知时，宏观特征占主要因素；在细粒度认知时，宏观特征和微观特征都很重要。本文数据集粒度有粗有细，因此q取1到⌈lb$N$⌉。在实际应用中，q的取值范围可以根据需要进行调节，最为典型的，如，水文分析领域或股票金融市场，只需要取q=2获得Hurst指数即可。

3.2　计算单条流的空时分形

步骤1　获得空间序列{P_i}和时间序列{T_i}。大多数流量抓包软件可提供每个数据包的大小以及到达时间信息。以QQ即时通话视频流为例，通过Wireshark抓包可获得

$$\{P_i\}=\{470,462,1494,\cdots,68,1494,1494\}$$

$$\{T_i\}=\{0.000428,0.00083,\cdots,0.151786,0.05897\}$$

步骤2　生成空时分形。对不同的观测尺度q=1,2,⋯,⌈lb$N$⌉，由式(15)~式(17)对时间序列和空间序列生成对应的分形特征$\boldsymbol{f}_P(\alpha)$和$\boldsymbol{f}_T(\alpha)$。

$$\boldsymbol{f}_P(\alpha)=\{16.513,13.436,10.237,7.288,4.362,$$
$$3.538,3.192,2.641,2.407,2.215\}$$

$$\boldsymbol{f}_T(\alpha)=\{6.285,5.217,4.163,3.722,2.338,1.176,$$
$$1.035,0.919,0.814,0.752\}$$

由式(18)生成空时分形$\boldsymbol{M}=\boldsymbol{f}_P(\alpha)*\boldsymbol{f}_T(\alpha)^{\mathrm{T}}$。如图9所示，空时分形在空间和时间2个维度上刻画出更多的流量细节特征点。

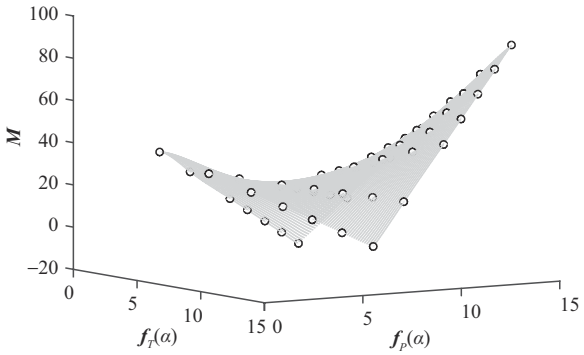


图9　空时分形特征点

相较于图5的传统分形，图10所示的空时分形可以清晰分辨SD和HD这2种类型流量。此外，图5的传统分形是基于10 000个数据包计算得到，而空时分形仅需1 500个数据包就可以获得稳定的特征，使SFM在确保准确率的基础上大幅提升计算速度。下文介绍SFM的认知精度和认知速度。

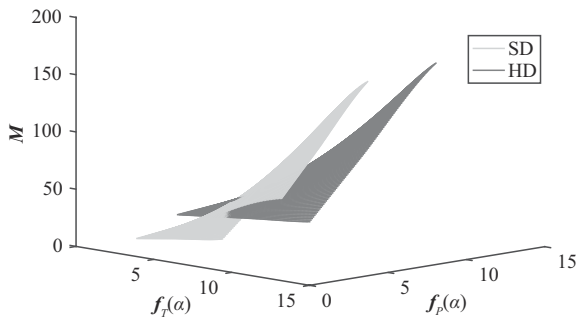


图10　优酷标清和高清视频流量的空时分形特征曲面

3.3　认知精度

本文采用分类系统常用的指标[27]：准确率、召回率和F值来评价分类准确率。在NJUPT数据集上，针对6类流量随机选择6 000条（每种类型1 000条）进行二折交叉验证，取20次分类结果的平均值，统计结果如表1所示。由表1可知，6类流量平均F值、准确率和召回率分别为95.86%、95.84%和95.88%。

表1　准确率统计

流量类型	F值	准确率	召回率
FTP传输	97.13%	97.35%	96.92%
电子邮件	95.56%	95.74%	95.39%
流媒体视频	95.86%	94.95%	96.78%
VoIP	96.82%	96.63%	97.02%
网页浏览	94.09%	95.12%	93.09%
网络游戏	95.69%	95.27%	96.12%

对多种方法，包括统计特征方法SDM[13]、行为特征方法BFM[14]、分形指数方法$D(h(q))$ DHQ[18]，以及分形谱方法FS[19]进行训练测试。具体的参数设置如表2所示，限于文章篇幅，这里仅列出最重要的参数，其他参数详见文献[13-14, 18-19]。分类结果如表3所示。现有方法对粗分类（NJUPT数据集六大类）均有良好表现，尤其是BFM，召回率高达99%。

表2　参数设置

对比方法	参数
SDM	隐层数量H_D=4，每层神经元数量为40，batch =1 000
BFM	权重系数ρ=0.8，迭代次数为100，ICR 和IPR的阈值σ=0.1、μ=0.3
DHQ	解析度N=10 000，阶层q=[−5,+5]，小波级数j=[3,6]
FS	解析度N=10 000，核阈Q=15，分段s=8
SFM	解析度N=1 500，观测尺度为ln$N$

在线流量数据具有以下特点：1) 随着万物互联技术发展，新应用不断产生，新类型流量不断涌现，目标类是变动的；2) 网络应用越来越多，流量类型越来越多，认知粒度越来越细。针对流量数据特点，用UNB数据集来验证各种方法面对变化目标类的表现。从表4可以看出，性能下降最快的是BFM，F值从98.45%下降到88.62%。基于时间关联的行为特征是研究人员针对特定类型流量积累的先验知识，面对变化目标类不能有效响应。SDM的F值也下降到91.62%，当目标类改变，一些统计特征变得不再有效，若要恢复模型的认知能力则需要针对新数据集重新使用特征选择算法筛选有效特征。随着目标类持续变动，统计特征方法受到更多限制。DHQ的认知准确率也有所下降，该方法基于分形指数$D(h(q))$对网络流实施分类，因缺乏细节特征导致性能受限。FS的连续分形谱极大改善了分形指数$D(h(q))$的不足，因此分类准确率较高，但连续分形谱的获取过程复杂，下文将论证这点。本文提出的空时分形在空间和时间2个维度刻画流量的变化特征，这些细节特征使SFM在面对变化目标类仍具有较强的认知准确率。

表3　粗粒度认知对比

对比方法	F值	准确率	召回率
SDM	94.48%	95.83%	93.17%
BFM	98.45%	97.82%	99.09%
DHQ	95.62%	97.35%	93.95%
FS	95.24%	93.60%	96.93%
SFM	96.09%	95.44%	96.76%

表4　变化目标类对比

对比方法	F值	准确率	召回率
SDM	91.62%	92.33%	90.92%
BFM	88.62%	89.25%	88.00%
DHQ	91.55%	91.52%	91.59%
FS	94.09%	95.76%	92.47%
SFM	93.48%	92.65%	94.32%

用ISP数据集进一步测试细粒度类别的认知性能，结果如图11所示。BFM的平均分类准确率下降至73%。行为特征以数据强相关为基础，因此，数据丢失会导致认知能力下降，关键性的握手、应答数据包丢失则会导致认知失败，行为特征在

非平稳网络环境下的泛化能力普遍较弱。SDM的统计特征对细粒度类别的认知效果也不佳，如对L_3的辨识度不到70%。当目标类增多，类型间差异减小，一些统计特征不再有效，导致性能下降。DHQ对有些流量无法识别，如对L_2的辨识度仅仅是62%。SFM与FS的认知性能相当，平均准确率达到92%，FS基于连续分形谱特征对流量进行精准细粒度认知；SFM基于空时分形进行细粒度认知，空时分形分别从时空维度观测流量的变化特征，比传统分形刻画更多细节特征，以此获得精准的细粒度认知能力。

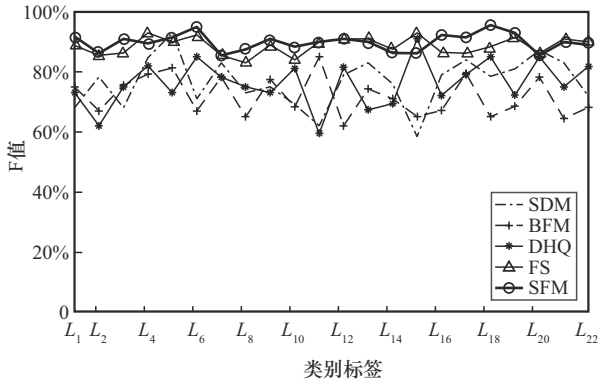


图11　细粒度认知(22类)

接下来，测试各种方法面对噪声环境的鲁棒性，在NJUPT数据集中随机修改数据包来模拟噪声，设置噪声比例分别为5%、10%和15%，实验结果如表5所示。BFM抗噪能力较差，关键性的握手、应答数据包被干扰导致认知失败。DHQ抗噪能力较强，$D(h(q))$分形指数反映宏观上的数据突发水平，因此一些数据包因噪声被篡改不会太影响认知结果。FS和SFM既包含了宏观上的数据突发水平，又体现细节上的数据突发水平，因此当某个数据包产生较大干扰时，局部细节特征的突变将导致认知失败。

表5　面对噪声数据的性能表现(F值)

对比方法	噪声比例为5%	噪声比例为10%	噪声比例为15%
SDM	90.35%	86.70%	79.78%
BFM	92.22%	86.04%	77.02%
DHQ	93.17%	89.27%	84.09%
FS	92.50%	88.73%	82.92%
SFM	92.69%	88.95%	83.12%

3.4　认知速度

流量认知不仅要保证较高的准确率，还要保证较低的时空复杂度，为此统计各种认知模型的训练时间和在线计算时间。如表6所示，SDM所需的训练时间最长。在SDM中，深度学习系统采用4层×40单元的设计，训练模型参数需经历400次迭代，训练神经网络非常耗时。BFM是基于朴素贝叶斯的行为特征方法，所使用的行为特征需要对网络流数据包进行遍历，因此BFM的训练时间主要集中在数据包遍历和概率计算，复杂度相对较低。基于分形的3种方法（DHQ、FS、SFM）所需的训练时间较少，基于分形指数的DHQ基于$D(h(q))$和$h(q)$ $(-5\leqslant q\leqslant -1)$总计10个分形特征用支持向量机（SVM）分类器对网络流实施分类，DHQ的训练时间主要是SVM分类器的训练时间，因此随着数据集中目标类的增多，训练时间呈递增趋势。FS使用连续分形谱进行分类，基于标记样本通过训练获得阈值T，其训练时间主要取决于训练样本的个数。SFM与FS类似，基于式(32)训练得到阈值T即可。

表6　训练时间

对比方法	训练时间/s		
	NJUPT	UNB	ISP
SDM	614.456	595.707	1 366.013
BFM	83.223	94. 935	257.619
DHQ	37. 195	43.450	129.262
FS	27.217	38.359	107.043
SFM	19.266	27.404	77.081

在线认知计算时间如图12所示，SDM所需的时间最多，4层×40单元的设计使得仅权重计算就多达6 400次。当面对不同数据集时，SDM的认知计算速率基本不变，因为深度学习系统的认知计算速率主要取决于所设计的网络架构而非数据集。在线分类的时空复杂度如表7所示，其中，D是训练样本数量，N是流序列解析度，J是统计特征数量，M是分类样本数量，K是特征值的数量，L是流的类别数量。本文不对SDM的时空复杂度进行理论分析，因为深度学习的复杂度涉及很多系统参数和超参数，如输入数据维度、隐藏层权

重维度、卷积核大小、激活函数等。BFM 的分类速率随着目标类的增多（NJUPT 数据集有 6 个目标类，UNB 数据集有 8 个目标类，ISP 数据集有 22 个目标类）而呈现递增趋势。随着目标类增多，BFM 所需行为特征增多，增加了计算量，BFM 的时空复杂度与统计特征的数量 L 和 K 成正比。DHQ 和 FS 的认知速率也随着类别数量呈现递增趋势，因为 DHQ 和 FS 的时空复杂度与目标类数量 L 呈正相关。

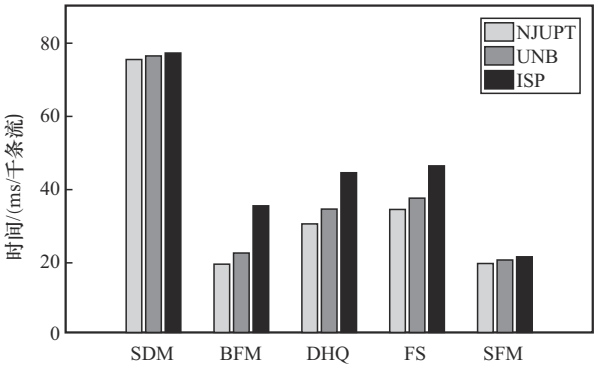


图 12　在线认知计算时间

表 7　在线分类的时空复杂度

方法	时间复杂度	空间复杂度
BFM	$O(MLK)$	$O((D+M)LK)$
DHQ	$O(ML^2N\,\mathrm{lb}N)$	$O(MLN\,\mathrm{lb}N)$
FS	$O(MLN\,\mathrm{lb}N)$	$O((M+L)N\,\mathrm{lb}N)$
SFM	$O(MN\mathrm{lb}N)$	$O((M+L)\,(\mathrm{lb}N)^2)$

当数据集变化时，SFM 分类速率基本保持恒定，处理每千条流量的分类仅需 19 ms 左右。SFM 在线分类的计算量主要集中在以下几个方面。1) 数据的预处理。时空分离后的时间序列和空间序列分别生成分形特征。由式(15)~式(17)可知，这一过程的计算量主要是扫描流量进行聚并求和，即 $O(N\mathrm{lb}N)$，N 是流序列解析度。2) 生成空时分形。取观测维度 $q=\lceil \mathrm{lb}\,N \rceil$，因此基于式(18)生成空时分形的计算量为 $O((\mathrm{lb}N)^2)$。3) 分类。这一过程主要是计算待测流量与各中心点的差异度 $\mathrm{Sim}(\boldsymbol{M}_{F_l},\boldsymbol{M}_{P_i})$，然后根据相似度来归类。因为本文中 $\mathrm{tr}(\boldsymbol{f}_P(\alpha)\boldsymbol{f}_T(\alpha)^{\mathrm{T}})=\boldsymbol{f}_T(\alpha)^{\mathrm{T}}\boldsymbol{f}_P(\alpha)$，于是式(29)的计算大大简化，观测维度 $=\lceil \mathrm{lb}\,N \rceil$，因此计算量为 $O(L\mathrm{lb}N)$，L 为类的数目。于是得到总的算法复杂度为 $O(N\mathrm{lb}N+(\mathrm{lb}N)^2+L\mathrm{lb}N)$。如果是 M 条流参与认知计算，则时间复杂度为 $O(MN\mathrm{lb}N)$。

此外，考察空间复杂度。将待测流量与 L 个类中心点进行比较并进行归类，计算所需要的存储空间主要取决于存储各个空时分形。本文取观测维度 $q=\lceil \mathrm{lb}\,N \rceil$，因此空时分形所需要的内存空间为 $O((\mathrm{lb}N)^2)$。L 个类中心点加上待测流量，因此空间复杂度为 $O((M+L)(\mathrm{lb}N)^2)$。

综上所述，SFM 的时间复杂度和空间复杂度都是比较小的，适合于在线的流量类型认知检测。传统分形（如 FS 分形谱）是空间维度和时间维度的融合特征，SFM 则基于空间和时间 2 个维度建立空时分形，刻画流量在空间维度和时间维度上更为细致的分形特征，这些细节特征使得 SFM 仅需较少的数据包就可以获得良好的认知性能，大幅提升了计算速率。

4　结束语

不同于统计特征和行为特征，分形特征以数据相关性为基础、以随机数据的变化特性为内容，适合于高可变的网络环境。本文最大贡献在于创新性提出空时分形，其物理含义是从不同空间、时间尺度观测流量并获得其变化特性。相比传统分形，空时分形描述更多细节特点，使 SFM 仅需 1 500 个数据包就可以实施高效认知，在降低模型复杂度的同时提升认知精度。本文对 22 种流量在保证 92% 的认知准确率的基础上，实现每千条流 19 ms 的认知速率，但是当前二分类方法已突破每千条流仅 10 ms 的高速认知。下一步研究将尝试改良分形特征结构，目标是仅需数百个包就可以进行早期预测，以此获得多类型流量的超高速认知。

参考文献：

[1] 张平, 张建华, 戚琦, 等. Ubiquitous-X: 构建未来 6G 网络[J]. 中国科学: 信息科学, 2020, 50(6): 913-930.
ZHANG P, ZHANG J H, QI Q, et al. Ubiquitous-X: constructing the future 6G networks[J]. Scientia Sinica (Informationis), 2020, 50(6): 913-930.

[2] 尤肖虎, 尹浩, 邬贺铨. 6G 与广域物联网[J]. 物联网学报, 2020, 4(1): 3-11.
YOU X H, YIN H, WU H Q. On 6G and wide-area IoT[J]. Chinese

Journal on Internet of Things, 2020, 4(1): 3-11.

[3] XIAO Y, XIA R, LI Y, et al. Distributed traffic synthesis and classification in edge networks: a federated self-supervised learning approach[J]. IEEE Transactions on Mobile Computing, 2024, 23(2): 1815-1829.

[4] SHEN M, YE K, LIU X, et al. Machine learning-powered encrypted network traffic analysis: a comprehensive survey[J]. IEEE Communications Surveys and Tutorials, 2023, 25(1): 791-824.

[5] 6G white paper on edge intelligence[R]. 2020.

[6] WU Z, DONG Y TIAN W, et al. Enhanced rough K-means based flow aggregation for QoS mapping in heterogeneous network environments[J]. IEEE Transactions on Network and Service Management, 2020, 17(2): 1197-1210.

[7] 6G: the next horizon, from connected people and things to connected intelligence (Huawei, whitepaper) [R]. 2021.

[8] GARCIA J, BRUNSTROM A. Clustering-based separation of media transfers in DPI-classified cellular video and VoIP traffic[C]//IEEE Wireless Communications and Networking Conference. Piscataway: IEEE Press, 2018: 1-6.

[9] YUN X, WANG Y, ZHANG Y, et al. A semantics-aware approach to the automated network protocol identification[J]. IEEE/ACM Transactions on Networking, 2016, 24(1): 583-595.

[10] 顾纯祥, 吴伟森, 石雅男, 等. 基于自编码器的未知协议分类方法[J]. 通信学报, 2020, 41(6): 88-97.
GU C X, WU W S, SHI Y N, et al. Method of unknown protocol classification based on autoencoder[J]. Journal on Communications, 2020, 41(6): 88-97.

[11] DONG S. Multi class SVM algorithm with active learning for network traffic classification[J]. Expert Systems with Applications, 2021, 176: 114885.

[12] 汤萍萍, 董育宁. 小波域基于分段Hurst指数的视频流分类[J]. 电子与信息学报, 2017, 39(6): 1298-1304.
TANG P P, DONG Y N. Classifying video flows based on segmented Hurst exponent in wavelet domain[J]. Journal of Electronics & Information Technology, 2017, 39(6): 1298-1304.

[13] WANG Z, MAO S, YANG W. Deep learning approach to multimedia traffic classification based on QoS characteristics[J]. IET Networks, 2019, 8(3): 145-154.

[14] WU Z, DONG Y N, WEI H L, et al. Consistency measure based simultaneous feature selection and instance purification for multimedia traffic classification[J]. Computer Networks, 2020, 173: 107190.

[15] HAJJAR A, KHALIFE J, DÍAZ-VERDEJO J. Network traffic application identification based on message size analysis[J]. Journal of Network and Computer Applications, 2015, 58: 130-143.

[16] WANG L, MEI H, SHENG V S. Multilevel identification and classification analysis of tor on mobile and PC platforms[J]. IEEE Transactions on Industrial Informatics, 2021, 17(2): 1079-1088.

[17] YANG L, 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.

[18] ARESTRÖM E, CARLSSON N. Early online classification of encrypted traffic streams using multi-fractal features[C]//EEE Conference on Computer Communications Workshops. Piscataway: IEEE Press, 2019: 84-89.

[19] TANG P P, DONG Y N, JIN J, et al. Fine-grained classification of Internet video traffic from QoS perspective using fractal spectrum[J]. IEEE Transactions on Multimedia, 2020, 22(10): 2579-2596.

[20] 曾凡一, 苘大鹏, 许晨, 等. 新增未知攻击场景下的工业互联网恶意流量识别方法[J]. 通信学报, 2024, 45(6): 75-86.
ZENG F Y, MAN D P, XU C, et al. Identification method for malicious traffic in industrial Internet under new unknown attack scenarios[J]. Journal on Communications, 2024, 45(6): 75-86.

[21] CHEN Z H, CHENG G, WEI Z J, et al. Classify traffic rather than flow: versatile multi-flow encrypted traffic classification with flow clustering[J]. IEEE Transactions on Network and Service Management, 2024, 21(2): 1446-1466.

[22] ZHANG J L, LI F H, YE F. Sustaining the high performance of AI-based network traffic classification models[J]. IEEE/ACM Transactions on Networking, 2023, 31(2): 816-827.

[23] DONG S, XIA Y. Network traffic identification in packet sampling environment[J]. Digital Communications and Networks, 2023, 9(4): 957-970.

[24] KORNYCKY J, ABDUL-HAMEED O, KONDOZ A, et al. Radio frequency traffic classification over WLAN[J]. IEEE/ACM Transactions on Networking, 2017, 25(1): 56-68.

[25] ZHAO R, ZHAN M W, DENG X W, et al. A novel self-supervised framework based on masked autoencoder for traffic classification[J]. IEEE/ACM Transactions on Networking, 2024, 32(3): 2012-2025.

[26] ZHANG X X, HAO L, GUI G, et al. An automatic and efficient malware traffic classification method for secure Internet of things[J]. IEEE Internet of Things Journal, 2024, 11(5): 8448-8458.

[27] LU M, ZHOU B, BU Z Y. Two-stage distillation-aware compressed models for traffic classification[J]. IEEE Internet of Things Journal, 2023, 10(16): 14152-14166.

[28] SHI H, LI H, ZHANG D, et al. An efficient feature generation approach based on deep learning and feature selection techniques for traffic classification[J]. Computer Networks, 2018, 132: 81-98.

[29] HOROWICZ E, SHAPIRA T, SHAVITT Y. Self-supervised traffic classification: flow embedding and few-shot solutions[J]. IEEE Transactions on Network and Service Management, 2024, 21(3): 3054-3067.

[30] KUMAR R, SWARNKAR M, SINGAL G, et al. IoT network traffic

classification using machine learning algorithms: an experimental analysis[J]. IEEE Internet of Things Journal, 2022, 9(2): 989-1008.

[31] LELAND W E, TAQQU M S, WILLINGER W, et al. On the self-similar nature of Ethernet traffic (extended version) [J]. IEEE/ACM Transactions on Networking, 1994, 2(1): 1-15.

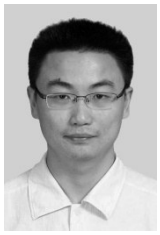
[32] WANG K, LI X, JI H, et al. Modeling and optimizing the LTE discontinuous reception mechanism under self-similar traffic[J]. IEEE Transactions on Vehicular Technology, 2016, 65(7): 5595-5610.

[33] STRELKOVSKAYA I V, GRYGORYEVA T I, SOLOVSKAYA I N. Self-similar traffic in G/M/1 queue defined by the Weibull distribution[J]. Radioelectronics and Communications Systems, 2018, 61(3): 128-134.

[34] UNB ISCX VPN-nonVPN traffic dataset[R]. 2024.

[作者简介]

汤萍萍（1981– ），女，安徽芜湖人，博士，安徽师范大学副教授、硕士生导师，主要研究方向为网络流分类传输、多媒体数据通信、QoS 保证技术、边缘环境与边缘智能等。



张晖（1982– ），男，山东邹城人，博士，南京邮电大学教授、博士生导师，主要研究方向为泛在异构网络流量分析、5G/6G 系统、智能终端与人工智能等。

董育宁（1955– ），男，江苏南京人，博士，南京邮电大学教授、博士生导师，主要研究方向为网络流识别与分类、无线网络、多媒体通信、无线通信网络等。

董国青（1986– ），女，山东聊城人，南京邮电大学博士生，安徽师范大学讲师，主要研究方向为网络流分类传输、无线通信网络、QoS 保证技术、物联网、5G/6G 系统、智能超表面技术等。