基于粒关系矩阵的流量在线分类

汤萍萍[1,2],董育宁[1]
(1. 南京邮电大学通信与信息工程学院,江苏南京 210003; 2. 安徽师范大学物理与电子信息学院,安徽芜湖 241000)

摘 要: 随着各种网络应用爆发式增长,流量的在线分类陷入困境之中. 传统的基于包统计特征的机器学习方法适用于稳定的网络环境,当网络拥塞出现严重的时延和丢包时将产生较大误差. 因而本文提出基于粒计算模型的分类方法. 粒计算属于人工智能计算的分支,当数据缺失、信息不完全或是有噪数据仍拥有较高的分辨能力. 为此本文将网络流量定义成粒子并构造粒子间关系,再建立粒关系矩阵. 传统的包统计特征只是粒关系矩阵当观测角度达到最大时的特例,因此粒关系矩阵对流量特性的描述更为全面,以此进行分类也更为精准. 最后实验数据证明了该方法的有效性和优越性.

关键词: 网络流量; 在线分类; 粒计算; 关系矩阵; 差异度
中图分类号: TN919 **文献标识码:** A **文章编号:** 0372-2112 (2021)01-0001-07
电子学报 URL: http://www.ejournal.org.cn **DOI:** 10.12263/DZXB.20191174

Online Traffic Classification Based on Granular Relation Matrix

TANG Ping-ping[1,2], DONG Yu-ning[1]
(1. *College of Telecommunications and Information Engineering, Nanjing University of Posts and Telecommunications, Nanjing, Jiangsu* 210003, *China*; 2. *College of Physics and Electronic Information, Anhui Normal University, Wuhu, Anhui* 241000, *China*)

Abstract: Online traffic classification is getting into troubles when the network applications are exploding. The traditional machine learning methods based on statistical characteristics of packets work well in stable network environment, but not in congestion environment with serious delay and packet loss. Therefore, a novel classification method based on granular computing is proposed in this paper. Granular computing belongs to the field of artificial intelligence computing, which is usually used to process missing, incomplete or noisy data. So we first define granules for the traffic, then construct the relations between the granules, and finally establish the relation matrix. The traditional statistical characteristics are only the special case of the relation matrix when the scale is the largest. The granular relation matrix can describe the traffic more comprehensively and classify them more accurately. The experiment results show its validity and advantages when compared with other methods.

Key words: network traffic; online classification; granular computing; relation matrix; difference degree

1 引言

随着网络技术的迅猛发展,网络应用呈爆发式增长,为此,研究者们提出一系列传输策略用以提高网络运行效率[1,2],如流量工程、容量规划、流量可视化、节能调度等. 然而,这些策略的基础是,首先要对网络流量进行准确的分类,可见流量的分类研究意义重大;此外,流量分类在网络安全、流量计费等领域也具有重要意义[3].

早期的流量在线分类,是将应用程序映射到特定的端口号[4],于是,流量类型(应用程序)与端口号一一对应. 然而,对于很多应用程序,其端口号是跳变的;端口也可以设置重用,多个应用程序可使用同一个端口号[5]. 因此,依靠端口号对流量进行分类,其准确度难以保证. 深度包检测则是一种相对精准的分类方法[6],它探测每个包的有效载荷,搜索特定的关键字以识别该流量所属的类型. 但是这种方法侵犯个人隐私,对于加密流量更是无计可施. 于是学者提出协议解析方法[7],通过协议的语义解析推理出流量所属的类型. 这种方法虽然没有侵犯个人隐私,但是仅靠协议,能够分

收稿日期:2019-10-14;修回日期:2020-07-20;责任编辑:孙瑶
基金项目:国家自然科学基金(No. 61271233);安徽省高校自然科学研究基金(No. KJ2019A0491)

类的粒度极其有限.况且,如果分类是放在网络层,则只能看到数据比特流,看不到数据传输的协议策略[8].

另外一些研究则是探寻流量的统计特征,结合机器学习方法进行细粒度分类.这些方法在离线状态下表现出色,如文献[9]选取数据包大小(PS)和包到达时间间隔(IPT)这两个统计特征,基于距离进行聚类:$\| x_i - x_j \|_2 = \sqrt{\sum_u |x_{iu} - x_{ju}|^2}$,对 HTTP、POP3、POP3S、EMULE、unknown 进行分类,准确度达到 80%.然而当应用于在线时逐渐暴露出以下局限性:

(1) 上述变量值 x_i, x_j 必须可知.若为未知那么此题将无解.当处理这一数据缺失问题时[10],现有的分类方法大多以均值(最大值或最小值)填充,分类性能也必然受到影响.

(2) 统计特征里的大多数特征并不适合在线分类.如包数量、平均包大小等,这些值必须要等到流结束时,才能确定下来;而对于实时的在线分类,这是非常严峻的制约因素[11].

为此,有些研究者提出基于子流进行分类[12].即,将一条流分割成若干子流,研究子流中所包含的重要特征.如文献[13]对子流特征进行分析,从而实现游戏和 VoIP 流量的在线分类.这种方法在一定程度上提高了分类的实时性,但是,其他问题又逐渐暴露出来:

(1) 分类粒度有限.如文献[14]靠子流中的几个特定包,用于识别 HTTP 视频流量.

(2) 分类准确率难以保证.尤其是网络状况不佳的时候,或恰巧子流中的这几个包出现丢包、重传、乱序等情况,分类的准确性将直线下降[15].

综上所述,长流的统计特征有利于分类的稳定性和准确性,可是不利于在线分类[16].子流的统计特征,虽然大幅提高了分类的实时性,但是也只能实现平稳网络环境下的粗分类,网络状况不佳时,分类结果不稳定[17].总之,基于统计特征的分类方法不能解决引言中我们提出的在线分类问题,因此,需要探寻其他路径来攻克这一难题.

Zadeh A L, Hobss J R, Lin T Y 等一批科学家重新挖掘、探索、分析人类的学习思维过程,得出一种全新的学习模型,它不同于机器学习、深度学习、逻辑推理,这种学习模型叫"粒计算"[18,19].它依据"数据之间所呈现的关系"进行推理分析,可屏蔽干扰噪声;可处理缺失的不完整数据[20,21].我们尝试将粒计算的相关原理应用于网络流量的分类,提出一种半监督的基于粒关系矩阵的分类算法 GrC (Traffic Classification with Granular Computing).该方法首先将网络流量定义成粒子,然后构造粒子间关系,再基于粒子关系建立关系矩阵,最后基于粒关系矩阵进行分类.因此本文主要贡献概括为:

(1) 首次提出流量粒子的概念.流量粒子是由比特流的数据包聚合而成,使得数据处理的对象不再是单独的数据包,而是聚合包,从而可以有效处理缺失数据、噪声数据等.

(2) 定义一种全新的流量特征——粒关系矩阵.粒关系矩阵反映的是不同空间、时间尺度所观察的流量的变化特性,体现粒子之间的时空关联性.粒关系矩阵突破长流统计特征不适合在线分类的限制,也改进了子流统计特征不适合细粒度分类的局限.

(3) 提出一种新型的差异度量方法——粒关系矩阵的差异度.粒关系矩阵间的差异,表明了流量所体现的变化轨迹的相似程度.我们从理论上论证了这一差异度用于流量分类的合理性和正确性;实验数据也表明,当面对高可变的网络流,拥有良好的分类稳定性和准确性.

2 粒计算分类模型 GrC

粒计算分类模型基本有三个步骤[22]:首先确定基本粒子.其次,分析基本粒子间的关联信息建立结构粒.最后依据这些关联信息进行推理或分类.

2.1 流量粒子

在定义流量粒子之前,首先给出流的准确描述:流量由流组成,或者说,众多流聚合形成流量.有的流是单向的;有的流是双向的,这些流的上下行特点往往截然不同,需要分别计算.因此我们引用 Barakat 对流(Flow)的定义[23],即五元组 < Src IP, Dest IP, Src Port, Dest Port, Protocol >.于是本文将流描述为一组满足 Barakat 定义的数据包:

$$F \triangleq \{(P_i, T_i) \mid_{i=1,2,\cdots,n}\} \tag{1}$$

这里,P_i 指的是第 i 个数据包的大小,T_i 指的是该数据包与前一个数据包的间隔时间,n 为流中包含的数据包的个数.

此外流 F 可分割成若干子流,第 m 个子流 $F^{(m)}$:

$$F^{(m)} \triangleq \{(P_i, T_i) \mid_{i=p+1,p+2,\cdots,p+n_m-1}\} \tag{2}$$

$$\text{s.t.} \quad p = \sum_{i=1}^{m-1} n_i \tag{3}$$

这里 n_m 是第 m 个子流中包含的数据包个数.

基于上述流及子流的定义,再参照 Chakraborty 和 Pal 的粒子构成思路[19],在本文中我们提出两种邻域粒,一是体量粒子:

$$N_v(x) = \bigcup_{k=i}^{j} P_i \in U \tag{4}$$

$$\text{s.t.} \quad |P_i - P_{i+1}| < Thr_v \tag{5}$$

由式(5)可见,邻域粒考查的是相互邻近的数据包,如果包大小的差异在给定的阈值 Thr_v 范围内,那么就融合成一个粒子 $N(\ )$.对流量进行这种操作后,得到

$N_v(x)\mid_{x=1,2,\cdots,X}$,$X$ 为体量粒子的大小.

另一种邻域粒是时量粒子,其形成规则为:

$$N_t(y)=\bigcup_{k=i}^{j}T_i\in U \tag{6}$$

$$\text{s.t.}\quad |T_i-T_{i+1}|<Thr_t \tag{7}$$

如果包间隔的差异在给定的阈值 Thr_t 范围内,那么就融合成一个粒子,即可得 $N_t(y)\mid_{y=1,2,\cdots,Y}$,$Y$ 为时量粒子的大小.

由上述定义可知,邻域粒 $N(\ )$ 中的成员无法事先确定,而是由临近包之间的相近程度来决定. 这样的粒子构成,使得计算模型对缺失数据变得不那么敏感,也能很好的去除噪声数据. 而这一点,正是粒计算的根基思想之一[21]. 按照粒计算"自底向上"的分类模型架构,一旦粒子形成、粒子层构造完毕,接下来就是构建粒子之间的关系形成结构粒.

2.2　结构粒

粒计算模型需要在不同层次、不同尺度(角度)研究粒子之间的固有关系. 而 Mandelbrot 也曾用尺度的思想研究事物的性状[22]. 设 $\{F(t)\}$ 为某区间上的随机过程,尺度为 ε 的测度 $\mu(\varepsilon)$ 若满足关系式:

$$\mu(\varepsilon)\propto\varepsilon^{-\alpha} \tag{8}$$

则 α 可以视为在 ε 尺度上所呈现的特征,被称为 Holder 指数,也叫做奇异性指数. 随后 α 被广泛应用,如矿井瓦斯涌出量预测、水文水资源的分类等[23].

由式(1)和式(2),网络流符合 Mandelbrot 对 $\{F(t)\mid_{t=i}\}$ 的定义,因此我们将依据式(8)来建立网络流量粒子之间的关系,形成结构粒:

$$\alpha\triangleq\frac{1}{m}\ln\tau_m \tag{9}$$

$$\text{s.t.}\quad \tau_m\triangleq\sum_{k=1}^{Z/m}\left|\sum_{i=1}^{m}\bar{N}_{Z/m}(m(k-1)+l)\right|^2 \tag{10}$$

这里,N 可以是 $N_v(x)$ 体量粒子,也可以是 $N_t(y)$ 时量粒子;$\bar{N}$ 是对邻域粒里的成员做平均;$Z=\{X,Y\}$ 是邻域粒的大小;m 表示的是观测尺度;最小的观测尺度 $m=1$,即每个邻域粒 $N(\ )$ 作为单独的粒子来处理;最大的尺度 $m=Z$,即所有邻域粒融合成为一个粒子,对应的正是统计特征里的"平均包大小",可见,统计特征是结构粒当观测角度达到最大时的特例. 或者可以理解为,统计特征是一种静态特征;而结构粒则是一种动态特征:当观测尺度 m 从 1 到 N 变化时,流量数据所呈现出的变化轨迹.

2.3　粒关系矩阵

如 2.1 节所示,提出了体量粒子和时量粒子这两种邻域粒;将这两种流量粒子带入式(9)、式(10)中,从而形成两种结构粒:体量粒子的结构粒 α_v 和时量粒子的结构粒 α_t. 前者描述的是流量包大小的变化特征;后者描述的是流量包在时间上的突发特征. 将这两个向量进行叉乘,其物理含义就是,在不同空间和时间尺度上网络流所体现的数据突发量的变化特征. 因此可得:

$$\boldsymbol{C}\mid_{X*Y}\triangleq\alpha_v*\alpha_t^{\mathrm{T}} \tag{11}$$

这里 α_v 是基于体量粒子 $N_v(x)\mid_{x=1,2,\cdots,X}$ 建立的结构粒,观测尺度最小为 $m=1$,最大为 $m=X$,因此有 X 个观测值. 同理 α_t 是时量粒子 $N_t(y)\mid_{y=1,2,\cdots,Y}$ 对应的结构粒,当尺度 m 从 1 到 Y 变化时得到 Y 个观测值. "$^{\mathrm{T}}$"是矩阵的转置. 因此粒关系矩阵 $\boldsymbol{C}$ 是一个 $X*Y$ 阶矩阵.

2.4　粒关系矩阵差异度

粒关系矩阵 $\boldsymbol{C}$ 描述的是随着观测尺度的变化流量的变化轨迹. 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. 于是我们基于矩阵 $\boldsymbol{C}$ 考量其差异度与相似性,实现对网络流的精准标定. 为此借鉴矩阵间关系度量方法[24]为粒关系矩阵 $\boldsymbol{C}$ 定义相似度量:

$$D(\boldsymbol{C}_a,\boldsymbol{C}_b)\triangleq\frac{\boldsymbol{C}_a\boldsymbol{C}_b^{\mathrm{T}}+\boldsymbol{C}_b\boldsymbol{C}_a^{\mathrm{T}}}{\boldsymbol{C}_a\boldsymbol{C}_a^{\mathrm{T}}+\boldsymbol{C}_b\boldsymbol{C}_b^{\mathrm{T}}} \tag{12}$$

$\boldsymbol{C}_a$ 表示流 F_a 的粒关系矩阵,$\boldsymbol{C}_b$ 是 F_b 的粒关系矩阵. 在此说明一下关于矩阵计算的维度选取问题. 设 $\boldsymbol{C}_a$ 为一个 X_a*Y_a 阶矩阵,$\boldsymbol{C}_b$ 为一个 X_b*Y_b 阶矩阵. 对这两者进行比较时需要站在相同的观测角度去分析,因此分别取 $\min(X_a,X_b)$ 和 $\min(Y_a,Y_b)$ 即可.

对于相似矩阵 $\boldsymbol{A}$ 和 $\boldsymbol{P}^{-1}\boldsymbol{AP}$,可知 $\mathrm{tr}(\boldsymbol{P}^{-1}\boldsymbol{AP})=\mathrm{tr}(\boldsymbol{PP}^{-1}\boldsymbol{A})=\mathrm{tr}(\boldsymbol{A})$,这里 $\mathrm{tr}(\ )$ 为矩阵的迹,即相似矩阵有相同的迹. 再者,由式(11)可见矩阵 $\boldsymbol{C}$ 是体量粒子的结构粒 $\alpha_v(x)$ 与时量粒子的结构粒 $\alpha_t(y)$ 的叉乘,因此 $\mathrm{tr}(\alpha_t(y)\alpha_v(x)^{\mathrm{T}})=\alpha_t(y)^{\mathrm{T}}\alpha_v(x)$,于是将式(12)相似度量的矢量矩阵转换成一个标量,并称之为差异度:

$$Dif(\boldsymbol{C}_a,\boldsymbol{C}_b)\triangleq 1-\frac{\mathrm{tr}(\boldsymbol{C}_a\boldsymbol{C}_b^{\mathrm{T}}+\boldsymbol{C}_b\boldsymbol{C}_a^{\mathrm{T}})}{\mathrm{tr}(\boldsymbol{C}_a\boldsymbol{C}_a^{\mathrm{T}}+\boldsymbol{C}_b\boldsymbol{C}_b^{\mathrm{T}})} \tag{13}$$

由式(13)可得 $Dif(\boldsymbol{C}_a,\boldsymbol{C}_b)=Dif(\boldsymbol{C}_b,\boldsymbol{C}_a)$. $Dif(\ )$ 在 0 到 1 之间;值越小说明两者间差异越小,相似度越高,极端情况下 $Dif(\boldsymbol{C}_a,\boldsymbol{C}_a)=0$,即两者之间无差异.

2.5　判别与阈值设定

假设当前有 L 个类 $\{M_l\}_{l=1,2,\cdots,L}$,每个类有若干条流 $\{\cdots,F_j,F_k,\cdots\}$,中心点记为 $\{P_l\}_{l=1,2,\cdots,L}$. $Dif(\ )$ 服从 0-1 上的均匀分布,因此中心点由下述公式确定:

$$P_l\triangleq\min_{F_k\in M_l}\left\{\max_{j\neq k,F_j\in M_l}Dif(C_k,C_j)\right\} \tag{14}$$

由式(14)可知,中心点 P_l 与类内其他点 $\{\cdots,F_j,F_k,\cdots\}$ 的差异度均为一个比较小的量,正以此点为类内中心点. 判断某条流 F_a 是否属于 M_l 时,计算该条流与中心点的差异度 $Dif(C_a,C_{Pl})$,如若差异小于或等于阈值,那么 F_a 属于类 M_l;如若差异大于阈值,那么 F_a 不属于类 M_l:

$$\mathrm{Be}(F_a,M_l)\triangleq\begin{cases}\in,\text{if}\{Dif(C_a,C_{Pl})\leqslant T\}\\ \notin,\text{if}\{Dif(C_a,C_{Pl})>T\}\end{cases} \tag{15}$$

本文流量分类模型采用半监督的学习方式:先基

于人工标注样本训练系统,然后加入未标记样本,通过式(15)进行判别分类. 当数据积累到一定量时,对阈值等系统参数进行调整. 这里阈值 T 的调节是影响整个系统性能的重要指标. 阈值过窄将使分辨率过高,不利于归类;阈值过宽将使分辨率过低,不利于识别. 由自适应阈值确定方法 OTSU[25],类间差异最大意味着错分概率最小,即 $\min(frr+far)$,这里 frr 为 False Rejection Rate 拒真率,far 为 False Acceptance Rate 认假率. 因此,我们以最大的类间差为基础,建立全局最优阈值调整机制:

$$T^* \triangleq \operatorname{argmax} \sum_{i \neq j} (Dif^2(t; M_i \leftrightarrow M_j)) \tag{16}$$

其中 $Dif(t; M_i \leftrightarrow M_j)$ 是阈值为 t 时 M_i 与 M_j 之间的类间差异.

2.6 算法复杂度

阈值只有在训练阶段才会被反复计算;在测试阶段,当测试数据积累到一定量时,才对阈值等系统参数以离线的方式进行调节. 因此在线分类的计算量主要集中在:

(1) 数据的预处理,也就是流量粒子形成. 由式(4)~(7)可见,这一过程只要扫描一遍流量,即 $O(N)$,N 是流序列解析度.

(2) 生成粒矩阵. 我们取观测尺度 $m = \lceil \log N \rceil$,因此二维矩阵的生成计算量为 $O((\log N)^2)$.

(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 $Dif(C_{Fi}, C_{Pl})$,然后根据差异度来归类. 因为本文中 $\mathrm{tr}(\alpha_t(y)\alpha_v(x)^{\mathrm{T}}) = \alpha_t(y)^{\mathrm{T}}\alpha_v(x)$,于是式(13)的计算大大简化,结构粒的维数 = 观测尺度 = $\lceil \log N \rceil$,因此计算量为 $O(L(\log N))$,L 为类的数目. 于是得到总的算法复杂度为 $O(N + (\log N)^2 + L(\log N)) \approx O(N)$. 如果是 M 条流参与分类,则时间复杂度为 $O(MN)$. 可见这个计算量是非常小的.

另一方面,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. 因此,计算所需要的存储空间主要在于存储各个粒关系矩阵. 我们取观测尺度 $m = \lceil \log N \rceil$,因此粒关系矩阵所需要的内存空间为 $O((\log N)^2)$. L 个类中心点加上待测流量,因此空间复杂度为 $O((L+1)(\log N)^2) \approx O(L(\log N)^2)$.

3 实验

实验的软件环境:用 Wireshark 软件捕捉实时业务流;在 Microsoft Visual Studio 平台上基于 C ++ 开发数据预处理程序,将流量数据处理成定义(1)的方式;并模拟高可变网络环境下的网络流量供后续实验所用;基于上述取得的数据,使用 MATLAB R2016a 仿真工具来验证 GrC 方法的有效性. 硬件配置环境为 Win10 professional (64bit/SP1), Intel (R) Core (TM) i7-7500U @ 2.70 GHz,8 GB 内存.

本实验使用的数据集分两类:一是在南京邮电大学校园网内获取的 NJUPT 数据集,考虑到网络动态变化特性,我们在不同时期捕获流量;另一个是因特网流量数据集 UNB ISCX Network Traffic[26],我们从其官网下载了 28G 的网络业务数据,包含了众多应用程序的流量数据,诸如 Vimeo, YouTube, ICQ, Skype, Facebook, Bittorrent 等等.

实验 1 测试分类效果

在 NJUPT 数据集上,针对即时通讯、P2P 单向视频、单机游戏、流媒体 HD 这四类流量,随机选择 4000 条(每种类型各 1000 条),进行二折交叉验证. 我们取其中一次分类结果如表 1 混淆矩阵所示. 其中,即时通讯流量被识别成即时通讯类、单向视频类、单机游戏类、流媒体 HD 类的数量依次是 521、16、14、15;而单向视频流量被识别成即时通讯类、单向视频类、单机游戏类、流媒体 HD 类的数量依次为 13、471、12、17.

表 1 混淆矩阵/条

	即时通讯	单向视频	单机游戏	流媒体 HD	识别率
即时通讯	521	16	14	15	92.05%
单向视频	13	471	12	17	91.81%
单机游戏	11	12	402	10	92.41%
流媒体 HD	15	14	19	438	90.12%

由表 1 可计算即时通讯类、单向视频类、单机游戏类、流媒体 HD 类的 frr 分别为 7.95%, 8.19%, 7.06%, 9.88%;即时通讯类、单向视频类、单机游戏类、流媒体 HD 类的 far 分别为 8.15%, 8.47%, 6.25%, 8.26%. 由此可见,式(16)基于 OTSU 的阈值设定方案,是一种全局优化方法,可有效避免局部最坏情况,对各类型的流量均能良好地分类.

重复上述实验,取 20 次分类结果的均值. 如表 2 第一行所示,统计数据显示本文方法对即时通讯类、单向视频类、单机游戏类、流媒体 HD 类的识别率依次是 92.17%、91.5%、92.44%、91.09%.

表 2 识别率统计/%

	即时通讯	单向视频	单机游戏	流媒体 HD
GrC	92.17	**91.5**	92.44	91.09
I-SVM[27]	**94.61**	78.49	85.33	79.84
Fractals[28]	91.35	90.76	90.47	91.6
K-L[13]	83.82	76.44	**93.76**	84.47
SFNN[11]	86.74	90.68	82.89	**93.64**
CPRF[4]	81.28	89.17	82.89	78.61

实验 2 与其他方法比较

在实验 1 的基础上,用上述数据对多种方法,包括

I-SVM[27],Fractals[28],K-L[13],SFNN[11],以及 CPRF[4]方法分别进行训练测试,并与本文的方法 GrC 做横向比较,结果如表 2 所示.

总的来说,大多数方法对于小数据集(流量类型少)都具有较好的分类识别能力;所以接下来,我们将针对大数据集(流量类型多),来进一步测试各种方法的分类性能. 这里,我们对来自于 NJUPT 和 UNB 数据集的 22 个类别的数据进行分类. 仍然是:每种类型的流量随机选择 1000 条,进行二折交叉验证;取 20 次分类结果的平均值. 图 1 中,横坐标是这 22 种类别的标号 $L_i, i=1,2,\cdots,22$,纵坐标是分类方法对流量类型的识别率.

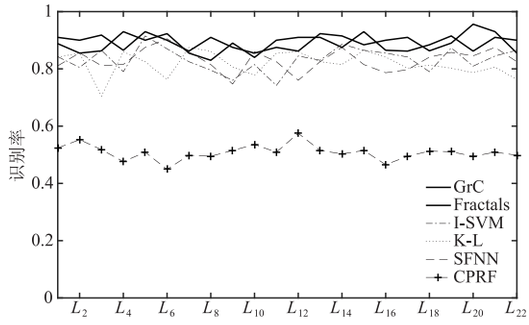

图1　大数据集测试

由图 1 可见,SFNN 和 K-L 以及 I-SVM 方法对某些流缺乏辨识度,如 K-L 对 L_3 的辨识度不到 70%,I-SVM 对 L_{12} 的辨识度也只有 72%;这些方法对某些类型的流又表现得相当敏感,如 I-SVM 对第 5 号流的辨识度达到 90% 以上. 此外,CPRF 的 CP 特征集对细粒度的视频流分类基本失效. 当视频种类增多,视频类别之间的差异更加微妙,CP 特征集已经无法识别和分辨.

在这个实验中,本文方法与 Fractals 方法的分类性能相当. Fractals 方法基于流量的分形特征,可以在细粒度上进行更为精准的分类. 而本文方法基于粒度计算模型,从时空观测角度观测流量的变化特征,以此分类,精准度也很高.

但是,本文研究的目标是在线分类. 在线分类面临着诸多不确定性,如网络拥塞、信号干扰、噪声数据等等. 因此,接下来,我们用动态数据进一步测试上述方法用于在线分类的真实性能.

实验 3　动态性能测试

为了模拟网络拥塞,我们对上述所有流量数据进行随机丢包、增加延时;为了模拟网络信号跳变、干扰等噪声数据,我们还对数据包进行篡改、增包. 在每条流内,这些修改的数据量控制在 5% 以内,变化幅度也设置在 5% 以内.

对上述流量数据重新进行分类识别,结果如图 2 所示,识别率都大打折扣,原因在于这些方法所训练好的有效特征总是基于平稳的、良好的网络环境,而当在线时,网络拥塞随时发生,严重程度也是非常随机的,一些影响识别性能的关键因素不可能进行实时更新,因此制约了在线识别的应用.

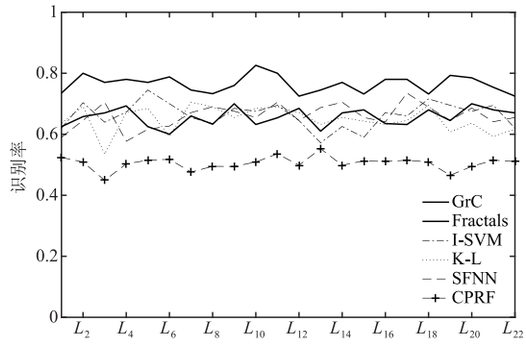

图2　动态性能测试

本文基于粒计算的流分类技术可有效屏蔽噪声、干扰数据,适合于高可变的动态网络;因此当网络环境变化或在线分类时具有较强的适应能力.

4　总结

本文深入探讨网络流量的细粒度在线分类问题,研究发现诸多统计特征并不适合在线分类;另外当网络面临拥塞、攻击等复杂状况下,我们需要寻找一种良好的抗干扰、能够屏蔽噪声数据的在线分类方法. 为此我们基于粒计算模型将网络流量的数据包进行聚合,形成粒子;并从时间空间角度研究粒子之间的关系,最后基于粒关系矩阵实现精准的分类. 文章中的一系列理论证明及实验也很好地表明了该方法在细粒度在线分类方面的有效性,以及与其他方法相比而体现的优越性.

此外,本文也存在若干问题需进一步探索:

(1)高维关系矩阵. 本文从时间、空间的不同观测角度建立了一个二维的粒关系矩阵. 我们将引入其他观测角度来建立一个高维的关系矩阵,以便于在更细粒度的分类上获得更高的精准度! 这是我们下一步要突破的研究难点,即超细粒度分类.

(2)“自顶向下”的粒计算分类模型. 本文创新地引入粒计算模型用于网络流量分类,但是跟大多数粒计算分类模型一致,采取的是“自底向上”的粒计算模型框架,即“基本粒子—结构粒子—分析推理”. 在未来我们将深入探索“自顶向下”的粒计算模型用于网络流量分类研究,期望有所突破和创新.

参考文献

[1] 谭晓衡,谢朝臣,郭坦. 基于区域感知贝叶斯决策的 5G

超密集异构网络联合垂直切换技术研究[J]. 电子学报, 2018,46(3):582 - 588.
TAN Xiao-heng, XIE Chao-chen, GUO Tan. Sensing Bayesian decision in ultra-dense HetNet for 5G[J]. Acta Electronica Sinica, 2018,46(3):582 - 588. (in Chinese)

[2] 黄建洋,兰巨龙,胡宇翔. 一种基于分段路由的多路径流传输机制[J]. 电子学报,2018,46(6):211 - 218.
HUANG Jian-yang, LAN Ju-long, HU Yu-xiang. A segment routing based multipath flow transmission mechanism[J]. Acta Electronica Sinica, 2018,46(6):211 - 218. (in Chinese)

[3] 申健,夏靖波,张晓燕,等. 基于分治排序策略的流量二次特征选择[J]. 电子学报,2017,45(1):128 - 134.
SHEN Jian, XIA Jing-bo, ZHANG Xiao-yan, et al. Secondary feature extraction of network traffic based on divide-conquer and ranking strategy[J]. Acta Electronica Sinica, 2017,45(1):128 - 134. (in Chinese)

[4] LI M, CHEN L H. Energy-efficient traffic regulation and scheduling for video streaming services over LTE-A networks[J]. IEEE Transactions on Mobile Computing, 2019, 18(2):334 - 347.

[5] GARCIA J, KORHONEN T, ANDERSSON R, VASTLUND F. Towards video flow classification at a million encrypted flows per second[A]. Proceedings of the International Conference on Advanced Information Networking & Applications [C]. Cracow, Poland: IEEE, 2018. 358 - 365.

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

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

[8] KORNYCKY J, ABDUL-HAMEED O, KONDOZ A, et al. Radio frequency traffic classification over WLAN[J]. IEEE Transactions on Parallel and Distributed Systems, 2017, 25(1):56 - 68.

[9] YAO L, QIN S, ZHU H. Feature selection algorithm for hierarchical text classification using Kullback-Leibler divergence[A]. International Conference on Cloud Computing & Big Data Analysis [C]. Chengdu, China: IEEE, 2017. 28 - 33.

[10] ORCZYKT AND PORWIK P. Investigation of the impact of missing value imputation methods on the k-NN classification accuracy[A]. Computational Collective Intelligence [M]. Berlin: Springer International Publishing, 2015. 557 - 565.

[11] SUN G, LIANG L, CHEN T, et al. Network traffic classification based on transfer learning[J]. Computers & Electrical Engineering, 2018, 25(3):177 - 193.

[12] CANOVAS A, JIMENEZ J M, ROMERO O, LLORET J, et al. Multimedia data flow traffic classification using intelligent models based on traffic patterns[J]. IEEE Network, 2018, 32(6):100 - 107.

[13] SHIM K S, HAM J H, SIJA B D. Application traffic classification using payload size sequence signature[J]. International Journal of Network Management, 2017, 27(5):e1981.

[14] GHOFRANI F, JAMSHIDI A, KESHAVARZ-HADDAD A, et al. Internet traffic classification using Hidden Naive Bayes model[A]. International Conference on Electrical Engineering[C]. Tehran, Iran: IEEE, 2015. 235 - 240.

[15] TEJERO-DE-PABLOS A, NAKASHIMA Y, SATO T. Summarization of user-generated sports video by using deep action recognition features[J]. IEEE Transactions on Multimedia, 2018, 20(8):2000 - 2011.

[16] LI W, YU X. An online flow-level packet classification method on multi-core network processor[A]. International Conference on Computational Intelligence and Security [C]. Shenzhen, China: IEEE, 2015. 407 - 411.

[17] AMBUSAIDI M, HE X, NANDA P, et al. Building an intrusion detection system using a filter-based feature selection algorithm [J]. IEEE Transactions on Computers, 2016, 65(10):2986 - 2998.

[18] FUJITA H, GAETA A, LOIA V, et al. Resilience analysis of critical infrastructures: a cognitive approach based on granular computing[J]. IEEE Transactions on Cybernetics, 2019, 49(5):1835 - 1848.

[19] CHAKRABORTY D B AND PAL S K. Neighborhood granules and rough rule-base in tracking [J]. Natural Computing, 2016, 15(3):359 - 370.

[20] BARAKAT C, THIRAN P, IANNACCONE G, et al. Modeling internet backbone traffic at the flow level[J]. IEEE Transactions on Signal Processing Special Issue on Networking, 2003, 51(8):2111 - 2124.

[21] ZADEH L A. Toward a generalized theory of uncertainty (GTU)-an outline[J]. Information Sciences, 2005, 172(16):1 - 2.

[22] MANDELBROT B B, WALLIS J R. Some long-run properties of geophysical records[J]. Water Resources Research, 2010, 5(2):321 - 340.

[23] HERNANDEZ-CARRASCO I, GARCON V, SUDRE J, et al. Increasing the resolution of ocean pCO2 maps in the south eastern Atlantic Ocean merging multifractal satellite-derived ocean variables[J]. IEEE Transactions on Geoscience & Remote Sensing, 2018, 56(11):2243 - 2249.

[24] DA K, YUN S, PARK H. SymNMF: Nonnegative low-rank approximation of a similarity matrix for graph cluste-

ring[J]. Journal of Global Optimization,2015,62(3):1 –30.

[25] OTSU N. A threshold selection method from gray-Level histogram[J]. IEEE Transactions on Systems Man & Cybernetics B,1979,9(1):94 –98.

[26] UNB ISCX VPN-nonVPN traffic dataset[OL]. http://www.unb.ca/cic/research/datasets/vpn.html.[2016].

[27] HAO S,HU J,LIU S,et al. Improved SVM method for internet traffic classification based on feature weight learning [A]. International Conference on Computer Applications and Information Sciences[C]. Changshu, China: IEEE, 2015. 102 –106.

[28] 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 Ping-ping,DONG Yu-ning. Classifying video fows based on segmented hurst exponent in wavelet domain [J]. Journal of Electronics & Information Technology, 2017,39(6):1298 –1304. (in Chinese)

作者简介

汤萍萍　女,1981 年生于安徽芜湖. 现为南京邮电大学通信与信息工程学院博士研究生,主要研究领域为多媒体数据通信、网络流分类传输、QoS 保证技术等.
E-mail:tpping@ahnu.edu.cn

董育宁(通信作者)　男,1955 年生于江苏南京. 博士、教授、博士生导师,主要研究领域为无线网络、多媒体通信、网络流分类等.
E-mail:dongyn@njupt.edu.cn