基于多任务特征学习的网络加密流量识别算法

孟　娟[1],孟　鹏[2],缪志敏[3],李晨溪[1],钱明远[1]
(1. 解放军 31108 部队,江苏 南京 210016;
2. 湖北科技学院,湖北 咸宁 437000;
3. 解放军陆军工程大学,江苏 南京 210007)

摘　要:加密数据流难以从其数据内容进行监管,但却是非法数据、敏感信息监管的重要对象。目前对加密数据流识别的研究大多依据特定的加密传输协议,主要通过端口匹配识别、深度包检测、深入流检测等来进行识别,这些方法实施的前提是加密协议已知,并未给出一种通用的加密数据流识别方法。对当前加密数据流识别技术进行了分析,分析加密数据流外在数据形式中所蕴含的内在属性信息,遵循“随机性特征——盲识别”的研究思路,研究一种通用的网络加密流量识别方法,利用加密流量的随机性特征,提出基于多任务特征学习的网络加密流量识别算法。该算法利用 $\ell_{2,1}$ 正则化项对一组相关任务进行联合特征学习。实验结果表明:该算法可有效识别网络加密流量,识别精度可达到 80% 以上。

关键词:加密流量识别;随机性;NIST 检验;特征选择;多任务特征学习
中图分类号:TP181　　**文献标识码**:A　　**文章编号**:1673-629X(2021)06-0112-06
doi:10.3969/j.issn.1673-629X.2021.06.020

Network Encrypted Traffic Identification Based on Multi-task Feature Learning

MENG Juan[1], MENG Peng[2], MIAO Zhi-min[3], LI Chen-xi[1], QIAN Ming-yuan[1]
(1. PLA 31108, Nanjing 210016, China;
2. Hubei University of Science and Technology, Xianning 437000, China;
3. Army Engineering University of PLA, Nanjing 210007, China)

Abstract: It's difficult to regulate the encrypted traffic from the content, but it is an important target to regulate the illegal data and sensitive information. Current researches on the encrypted traffic identification are more for specific encryption protocol, mainly through port information, load characteristics and flow characteristics. The implementation premise of these methods is that the encryption protocol is known. There is not a general method for the encrypted traffic identification. The technology of encrypted traffic identification is analyzed, followed the "randomness feature-protocol independent identification" research idea by analyzing the encrypted traffic inherent attribute information, and a general method of encrypted traffic identification is studied. Utilizing the randomness characteristics, a multi-task feature learning formulation is proposed to identify encrypted traffic, which captures the intrinsic relatedness among different tasks by a $\ell_{2,1}$-norm regularized multi-task feature learning model. Experiment shows that the identification accuracy of the proposed algorithm can get above 80% for encrypted traffic identification.

Key words: encrypted traffic identification; randomness; NIST test; feature selection; multi-task feature learning

0　引　言

随着互联网加密协议应用越来越广泛,网络加密流量的识别问题日益引起人们的重视。网络加密流量的有效识别,对保护用户信息,监管非法数据,检测网络攻击,维护网络安全有着重要意义。

尽管网络加密流量识别问题属于网络流量识别的子问题,但是传统的网络流量识别方法,如基于端口匹配识别、深度包检测、深入流检测等难以直接应用于网络加密流量识别中。目前对网络加密流量识别问题的研究多针对特定的加密协议,利用加密协议建立连接阶段的明文特征,通过特征码匹配来完成识别,或者利用加密协议建立连接阶段的报文指纹特征,如特定的报文长度、到达时间等来完成识别[1-7]。这些识别方法均针对某种特定的加密协议,并未给出一种通用的

收稿日期:2020-06-17　　**修回日期**:2020-10-20
基金项目:江苏省自然科学基金青年基金面上资助项目(BK20140075)
作者简介:孟　娟(1978-),女,博士,工程师,研究方向为模式识别、网络安全。

网络加密流量识别方法。

尽管网络流量的具体加密协议细节各不相同,但好的加密算法的一个最基本的要求是要保证算法是安全的。算法的安全性体现在算法数学基础的稳健性、算法的抗攻击性、算法的相对强度和算法输出序列的随机性。其中前 3 项因算法结构的种类而异,而利用统计检测原理来检测算法输出序列是否随机,只关心输入输出,在这种情况下,算法的具体结构是被忽略的,因此,可以通过网络流量的随机性特征来识别网络加密流量。

利用网络加密流量的随机性特征,该文提出了一种基于多任务特征学习的网络加密流量识别算法。该算法采用随机性检验获取数据流的随机性特征,提取多维随机性特征值,然后基于 $\ell_{2,1}$ 范式最小化对一组相关任务进行联合特征学习。最后针对网络加密流量进行实验分析,结果表明,该算法对网络加密流量的识别精度可达到 80% 以上。

1　网络加密流量的随机性特征

1.1　随机性度量

随机性度量是指通过判断被测序列是否存在某些随机序列的特定特征,进而判断其是否随机。随机性度量中需要对两个问题进行研究:第一个问题是如何检验,就是确定应该对被测序列的哪些特征进行统计分析,如频数、游程、相关性、累积和等,在设计随机性检验方法时,应该选择那些能反映随机特性的方法,并尽可能反映更多特性,这样通过较少的指标就可以全面地评价其随机性。第二个问题是如何对测试结果做出评估,就是用什么方法来对统计检验得到的结果进行准确的评价,评价方法有门限值、P-value 等。评价方法要尽可能准确。

现已知的随机性检验方法达 200 多种。其中具有代表性的是美国商务部国家标准技术协会(NIST)在 2001 年 5 月公布的 FIPS140-2 标准[8]和 SP800-22 标准[9]、德国资讯安全联合办公室(BSI)在 2001 年 9 月公布的 AIS31 标准[10]。国内也有随机性检验方法的相关研究[11-12]。

让一个随机序列通过所有的随机性检验,在时间、空间上需要很大的开支,同时有些检验方法反映的是随机序列同一方面的特性。对于网络加密流量而言,采用具有权威代表性的 NIST SP800-22 标准中的 15 种检验方法更加方便和准确,15 种检验方法及其所针对的序列属性简列如下:

(1)单比特频数检验。序列中 1 的个数。

(2)分组组内频数检验。对序列分组后每个子序列中 1 的个数。

(3)游程检验。序列中的游程个数。

(4)组内最长游程检验。对序列分组后每个子序列的最长游程长度。

(5)二进制矩阵秩检验。将序列构造成 N 个矩阵,计算每个矩阵的秩。

(6)离散傅里叶变换检验。序列进行傅里叶变换后的频率幅值。

(7)非重叠模式匹配检验。某特定模式的出现次数。

(8)重叠模式匹配检验。某特定模式的出现次数。

(9)Maurer 通用统计检验。特定长度子序列的所有模式中,相同模式间间隔距离的位数。

(10)线形复杂度检验。用 Berlekamp-Massey 算法计算每个子序列的线形复杂度。

(11)串行检验。特定长度模式出现的频数;

(12)近似熵检验。特定长度模式出现频率的熵值。

(13)累加和检验。序列累加和的最大值。

(14)随机游走检验。随机游走中各循环到达距离原点特定长度位置的次数。

(15)随机游走变体检验。随机游走中到达距离原点特定长度的位置的总次数。

该标准建立在假设检验的基础上,统一采用 P-value 评价方式。每个随机性检验的核心是其使用的统计量和统计量所满足的分布,每个统计量针对了一个特定的序列属性。

1.2　随机性特征的相关性

利用 NIST SP800-22 随机性检验可构建加密数据流的 15 类随机性特征,根据不同的参数设置可提取不同维度的随机性特征。按缺省设置运行 NIST 后得到统计结果,从统计结果中提取 188 维随机性特征值,其中单比特频数特征、分组频数特征关注的是序列中 0 和 1 出现的频率,游程特征、组内最长游程特征关注的是序列中比特变化的特点,离散傅里叶变换特征、非重叠模式匹配特征、重叠模式匹配特征、串行特征和近似熵特征关注的是序列中是否有某一个模式出现的频率较高。Maurer 通用统计特征和线性复杂度特征关注的是序列的压缩性。累加和特征、随机游动特征和随机游动变体特征关注的是序列的随机游动特性。各个统计特征都各有其关注点,从统计原理上来看,这些特征之间具有一定的联系,但是,从统计原理上分析特征之间相关性并不容易,有时候两个看起来差别很大的两个特征也有可能具有较高的相关性,这里通过计算它们的互信息熵对这些随机性特征进行相关性分析。

根据信息学中熵和互信息熵的意义,利用互信息熵来定量地衡量随机性特征相互之间的相关程度。每个随机性特征对应了一个统计量,当待测序列长度趋于无穷时,这个统计量近似服从于正态分布或卡方分布,但当待测序列长度一定时,这个统计量是一个离散的概率分布,根据信息熵的定义可以得到其熵值,即 $H(X)=-\sum_{i=1}^{n}p_i\log(p_i)$,p_i 为统计量每个可能取值出现的概率。熵值描述了该统计量值的不确定性。考虑条件熵:

$$H(X/b_j)=H_n(P(a_1,b_j),P(a_2,b_j),\cdots,P(a_n,b_j))=-\sum_{i=1}^{n}P(a_i/b_j)\log P(a_i/b_j)=-\sum_{i=1}^{n}\left[\frac{P(a_i,b_j)}{P(b_j)}\right]\log\left[\frac{P(a_i,b_j)}{P(b_j)}\right] \quad (1)$$

而

$$H(X/Y)=\sum_{j=1}^{s}P(b_j)H(X/b_j)=-\sum_{i=1}^{n}\sum_{j=1}^{s}P(a_i,b_j)\log P(a_i/b_j)=-\sum_{i=1}^{n}\sum_{j=1}^{s}P(a_i,b_j)\log\left[\frac{P(a_i,b_j)}{P(b_j)}\right] \quad (2)$$

由上式可得:

$$\begin{aligned}H(X/Y)&=-\sum_{i=1}^{n}\sum_{j=1}^{s}P(a_i,b_j)\log\left[\frac{P(a_i,b_j)}{P(b_j)}\right]=\\&-\sum_{i=1}^{n}\sum_{j=1}^{s}P(a_i,b_j)\log[P(a_i,b_j)-P(b_j)]=\\&-\left\{\sum_{i=1}^{n}\sum_{j=1}^{s}P(a_i,b_j)\log P(a_i,b_j)-\sum_{i=1}^{n}\sum_{j=1}^{s}P(a_i,b_j)\log P(b_j)\right\}=\\&H(X,Y)-H(Y)\end{aligned} \quad (3)$$

$H(X,Y)$ 是联合熵,$H(X,Y)$ 反映了 X 和 Y 这两个统计量信息量的总和:

$$H(X,Y)=-\sum_{i=1}^{n}\sum_{j=1}^{s}P(a_i,b_j)\log P(a_i,b_j) \quad (4)$$

$H(X)$ 是 X 的信息熵,$H(X/Y)$ 是给定 Y 的条件下 X 的信息量,根据式(3)和式(4)可以计算出统计量间的互信息熵。

$$\begin{aligned}I(X,Y)&=H(X)-H(X/Y)=\\&H(X)+H(Y)-H(X,Y)=\\&I(Y,X)\end{aligned} \quad (5)$$

互信息熵反映了两个统计量之间信息量的相互影响,可利用概率分布来估计互信息熵:

$$I(X,Y)=\sum_{x,y}P_{X,Y}(x,y)\cdot\log_2\frac{P_{X,Y}(x,y)}{P_X(x)\cdot P_Y(y)} \quad (6)$$

其中,$P_{X,Y}(x,y)$ 指 X 和 Y 的联合概率密度分布,$P_X(x)\cdot P_Y(y)$ 指 X 和 Y 各自独立时的概率分布。

互信息熵反映了两个统计量间的相互影响程度。当互信息熵为零时,这两个统计量相互独立,当互信息熵增大时,两个统计量间的相关程度增强。

2 多任务特征学习概率模型

多任务特征学习[13-19]旨在学习多个相关任务的共同特征表示。在网络加密流量识别中,将不同加密协议的网络加密流量识别看作不同的任务,尽管各个任务中网络流量的加密协议不同,但加密流量都具有随机性的特征,因此,可以通过多任务特征学习对多个任务的联合特征进行学习,识别网络加密流量。

在多任务特征学习中,对于 t 个任务 $\{(a_i^j,y_i^j)\}_{i=1}^{n_j},j=1,2,\cdots,t$,$a_i^j\in\mathbb{R}^d$ 表示第 j 个任务的第 i 个样本,y_i^j 表示对应样本的标签,n_j 表示第 j 个任务的训练样本数,$n=\sum_{j=1}^{t}n_j$ 表示总的训练样本数,$\boldsymbol{A}_j=[a_1^j,a_2^j,\cdots,a_{n_j}^j]^{\mathrm{T}}\in\mathbb{R}^{n_j\times d}$ 表示第 j 个任务的数据矩阵,$\boldsymbol{A}=[\boldsymbol{A}_1^{\mathrm{T}},\boldsymbol{A}_2^{\mathrm{T}},\cdots\boldsymbol{A}_t^{\mathrm{T}}]^{\mathrm{T}}\in\mathbb{R}^{n\times d}$,$\boldsymbol{y}_j=[y_1^j,y_2^j,\cdots,y_{n_j}^j]^{\mathrm{T}}\in\mathbb{R}^{n_j}$ 并且 $\boldsymbol{y}=[\boldsymbol{y}_1^{\mathrm{T}},\boldsymbol{y}_2^{\mathrm{T}},\cdots,\boldsymbol{y}_t^{\mathrm{T}}]^{\mathrm{T}}\in\mathbb{R}^n$,考虑线性模型:

$$f_j(a)=\boldsymbol{w}_j^{\mathrm{T}}\boldsymbol{a},\ j=1,2,\cdots,t \quad (7)$$

其中,$w_j\in\mathbb{R}^d$ 是第 j 个任务的权重向量,全部 t 个任务的权重向量形成权重矩阵 $\boldsymbol{W}=[\boldsymbol{w}_1,\boldsymbol{w}_2,\cdots,\boldsymbol{w}_t]\in\mathbb{R}^{d\times t}$ 。

假定第 j 个任务样本 $\boldsymbol{a}^j\in\mathbb{R}^d$,其对应的标签 $\boldsymbol{y}^j\in\mathbb{R}$ 符合均值为 $f_j(a^j)$,标准差为 $\boldsymbol{\sigma}^j(\boldsymbol{\sigma}^j>0)$ 的高斯分布:

$$p(y^j\mid w_j,a^j,\sigma^j)=\sqrt{\frac{\sigma^j}{2\pi}}\exp\left[\frac{-\sigma^j(y^j-w_j^{\mathrm{T}}a^j)^2}{2}\right] \quad (8)$$

记 $\boldsymbol{\sigma}=[\sigma_1,\sigma_2,\cdots,\sigma_t]^{\mathrm{T}}\in\mathbb{R}^t$,假定数据 $\{A,y\}$ 独立于分布式(8),则似然函数可以写成:

$$p(\boldsymbol{y}\mid\boldsymbol{W},\boldsymbol{A},\boldsymbol{\sigma})=\prod_{j=1}^{t}\prod_{i=1}^{n_j}p(\boldsymbol{y}_i^j\mid\boldsymbol{w}_j,\boldsymbol{a}_i^j,\boldsymbol{\sigma}^j) \quad (9)$$

为了捕获任务间的相关性,定义先验 $\boldsymbol{W}$,$\boldsymbol{W}$ 的第 i 行记为 $\boldsymbol{w}^i\in\mathbb{R}^{1\times t}$,对应所有任务的第 i 个特征,假设 $\boldsymbol{w}^i$ 由以下指数先验产生:

$$p(\boldsymbol{w}^i\mid\delta^i)\propto\exp(-\|\boldsymbol{w}^i\|\delta^i),i=1,2,\cdots,d \quad (10)$$

其中,$\delta^i>0$ 为超参数,当 $t=1$ 时,式(10)为拉普拉斯先验。假定 $\boldsymbol{\delta}=[\boldsymbol{\delta}^1,\boldsymbol{\delta}^2,\cdots,\boldsymbol{\delta}^d]^{\mathrm{T}}\in\mathbb{R}^d$,$\boldsymbol{w}^1,\boldsymbol{w}^2,\cdots,\boldsymbol{w}^d$ 独立于先验(10),$\boldsymbol{W}$ 的先验表示为:

$$p(\boldsymbol{W}\mid\boldsymbol{\delta})=\prod_{i=1}^{d}p(\boldsymbol{w}^i\mid\boldsymbol{\delta}^i) \quad (11)$$

W 的后验正比于先验和似然函数的乘积,即:

$$p(\boldsymbol{W}\mid\boldsymbol{A},\boldsymbol{y},\boldsymbol{\sigma},\boldsymbol{\delta})\propto p(\boldsymbol{y}\mid\boldsymbol{W},\boldsymbol{A},\boldsymbol{\sigma})\boldsymbol{p}(\boldsymbol{W}\mid\boldsymbol{\delta}) \quad (12)$$

取式(12)的负对数并结合式(7)~式(11),通过最小化式(13)可得到 $\boldsymbol{W}$ 的最大后验估计。

$$\frac{1}{2}\sum_{j=1}^{t}\boldsymbol{\sigma}^{j}\ \|\boldsymbol{y}_j-\boldsymbol{A}_j\boldsymbol{w}_j\|^2+\sum_{i=1}^{d}\boldsymbol{\delta}^{i}\|\boldsymbol{w}^{i}\| \tag{13}$$

简单起见,假定 $\boldsymbol{\sigma}=\boldsymbol{\sigma}^{j},\forall j=1,2,\cdots,t$,并且 $\boldsymbol{\delta}=\boldsymbol{\delta}^{i},\forall i=1,2,\cdots,d$,由式(13)可得到 $\ell_{2,1}$ 范式正则化的最小二乘回归问题,即:

$$\min\frac{1}{2}\sum_{j=1}^{t}\ \|\boldsymbol{y}_j-\boldsymbol{A}_j\boldsymbol{w}_j\|^2+\rho\ \|\boldsymbol{W}\|_{2,1} \tag{14}$$

其中,$\rho=\delta/\sigma$;$\|\boldsymbol{W}\|_{2,1}=\sum_{i=1}^{d}\|\boldsymbol{w}^{i}\|$ 为矩阵 $\boldsymbol{W}$ 的 $\ell_{2,1}$ 范式。

式(14)可泛化为以下 $\ell_{2,1}$ 范式正则化问题:

$$\min_{\boldsymbol{W}\in\mathbb{R}^{d\times t}}\text{loss}(\boldsymbol{W})+\rho\ \|\boldsymbol{W}\|_{2,1} \tag{15}$$

其中,$\rho>0$ 为正则化参数;$\text{loss}(\boldsymbol{W})$ 为凸光滑的损失函数,如最小二乘损失或 logistic 损失。

如果是单任务,则式(10)为拉普拉斯先验分布,此时式(15)就是 ℓ_1 范式正则化优化问题。如果是多个任务,第 i 个特征的权重通过 $\boldsymbol{w}^{i}$ 的二范式分组聚集,因此,$\ell_{2,1}$ 范式正则化倾向于对多个任务进行联合特征学习。

3　基于多任务特征学习的网络加密流量识别算法

利用网络加密流量的随机性特征,提出基于多任务特征学习的网络加密流量识别算法。首先采集网络数据流样本,对数据流进行预处理,获取载荷信息作为有效数据;然后,对有效数据进行 NIST SP800-22 随机性检验,获取数据流的多维随机性特征;最后对这些随机性特征进行多任务特征学习,利用学习模型对未知样本进行加密流量识别。算法流程如图 1 所示。

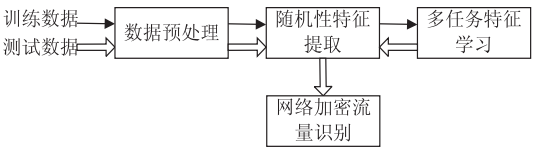

图 1　基于多任务特征学习的网络加密流量识别

3.1　数据预处理

对获得的网络数据进行预处理,获取有效数据,主要包括以下几个步骤:首先,根据传输链路的类型判断具体的链路层协议,丢弃无关内容后对网络层数据进行提取;然后,根据网络层数据协议类型分类处理,丢弃非 IP 协议报文,对 IP 协议报文去除报头内容后,将具有相同源、目的地址,相同源、目的端口以及相同协议的报文组成一个数据流;最后,对有效数据进行提取。如果数据流中的报文载荷是 TCP 或 UDP 协议报文,则去除相应的协议报头后剩余的数据即是提取的有效数据;如果数据流中的载荷不是 TCP 或 UDP 协议报文,则可直接将数据流中的报文载荷作为有效数据。

3.2　随机性特征提取

对数据预处理后得到的有效数据进行 NIST SP800-22 随机性检验,利用 NIST SP800-22 的 15 种检验方法,一共可提取 15 类随机性特征,根据不同的参数设置可得到不同维度的随机性特征值。该文按照缺省设置,得到 188 维随机性特征值。

3.3　多任务特征学习

将不同加密协议的网络加密流量识别看作不同任务,对提取出的多任务样本的随机性特征进行联合特征学习。对于 t 个任务样本 $\{(\boldsymbol{a}_i^j,\boldsymbol{y}_i^j)\}_{i=1}^{n_j},j=1,2\cdots,t$,每个样本具有 d 维(这里 $d=188$)随机性特征 $\boldsymbol{a}_i^j\in\mathbb{R}^d$,$\boldsymbol{y}_i^j$ 是其标签。在加密流量识别中,考虑分类模型 $f_j(x)=\text{sign}(\boldsymbol{w}_j^{\mathrm{T}}\boldsymbol{a}+c_j)$,其中 $\boldsymbol{w}_j$ 为该模型的权重向量,c_j 为模型偏差。$\boldsymbol{A}_j=[\boldsymbol{a}_1^j,\boldsymbol{a}_2^j,\cdots,\boldsymbol{a}_{n_j}^j]^{\mathrm{T}}\in\mathbb{R}^{n_j\times d}$,$\boldsymbol{A}=[\boldsymbol{A}_1^{\mathrm{T}},\boldsymbol{A}_2^{\mathrm{T}},\cdots,\boldsymbol{A}_t^{\mathrm{T}}]^{\mathrm{T}}\in\mathbb{R}^{n\times d}$ 为样本随机性特征的数据矩阵,$\boldsymbol{y}_j=[\boldsymbol{y}_1^j,\boldsymbol{y}_2^j,\cdots,\boldsymbol{y}_{n_j}^j]^{\mathrm{T}}\in\mathbb{R}^{n_j}$,$\boldsymbol{y}=[\boldsymbol{y}_1^{\mathrm{T}},\boldsymbol{y}_2^{\mathrm{T}},\cdots,\boldsymbol{y}_t^{\mathrm{T}}]^{\mathrm{T}}\in\mathbb{R}^{n}$ 为样本标记值矩阵,$\boldsymbol{W}=[\boldsymbol{w}_1,\boldsymbol{w}_2,\cdots,\boldsymbol{w}_t]\in\mathbb{R}^{d\times t}$ 为权重矩阵。

通过最小化如下目标函数来求解 $\boldsymbol{W}$:

$$\min_{W,c}\sum_{j=1}^{t}\sum_{i=1}^{n_j}\log(1+\exp(-\boldsymbol{y}_i^j(\boldsymbol{w}_j^{\mathrm{T}}\boldsymbol{a}_i^j+c_j)))+\rho\ \|\boldsymbol{W}\|_{2,1} \tag{16}$$

其中,$\boldsymbol{a}_i^j$ 表示第 j 个任务的第 i 个样本,$\boldsymbol{y}_i^j$ 为其对应的样本标签;$\boldsymbol{w}_j$ 为第 j 个任务的模型;c_j 为第 j 个任务的模型偏差;ρ 为正则化参数,控制随机性特征的组稀疏性。

利用加速梯度方法(accelerated gradient methods,AGM)[20]来求解此优化问题。AGM 不同于传统的梯度方法,每次迭代使用前两个点的线性组合作为搜索点,而不是仅使用最新点。AGM 收敛速度为 $O(1/k^2)$,这是最优的一阶方法,AGM 的关键子程序是计算邻近算子。

$$W^{*}=\arg\min_{W}M_{\gamma,S}(W)=\arg\min_{W}\frac{\gamma}{2}\|W-(S-\frac{1}{\gamma}\nabla L(W))\|_F^2+\Omega(W) \tag{17}$$

式中,$\Omega(W,\lambda)$ 为非平滑正则化项;γ 为步长,$\nabla L(\cdot)$ 为 $L(\cdot)$ 的梯度;S 为当前搜索点。

3.4　网络加密流量识别

对于测试样本 X,依据式(18)判断其是否为加密数据流:

$$y=\text{sign}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{W}+c) \tag{18}$$

4　实验与结果分析

首先采集不同加密协议的网络加密数据流和非加

密数据流。作为研究对象的加密数据流取自网上银行网站数据、VPN 加密通信数据、TOR 匿名通信和私有加密协议数据,非加密数据流取自普通新闻类网站数据和在线视音频数据,对网络流量进行预处理后得到有效载荷,对每一类数据,取 5 000 个样本,每个数据样本长度为 1 000 比特。然后对流量数据样本进行 NIST SP800-22 随机性检验,测试中的参数均使用缺省值,每个样本可得到 188 维随机性特征值。将得到的实验数据集划分为 50% 训练数据集和 50% 验证数据集。

4.1 识别参数选择

基于多任务特征学习的网络加密流量识别算法有一个 $\ell_{2,1}$ 正则化参数 ρ,在集合 $\{0.1, 0.05, 0.04, 0.03, 0.02, 0.01, 0.005, 0.001\}$ 中选择 ρ,学习的联合随机性特征数随参数 ρ 变化如表 1 所示(Fea 表示特征数)。随着 ρ 的减小,学习的联合随机性特征数逐渐增大。随着学习的联合随机性特征数的增大,获得的识别精度逐渐提高,如图 2 所示,在学习的联合随机性特征数达到 29 时识别精度趋于稳定,对应的 $\ell_{2,1}$ 正则化参数 $\rho = 0.02$。

表 1 学习的联合随机性特征数随参数 ρ 变化

ρ	Fea
0.1	4
0.05	7
0.04	12
0.03	22
0.02	29
0.01	65
0.005	86
0.001	159

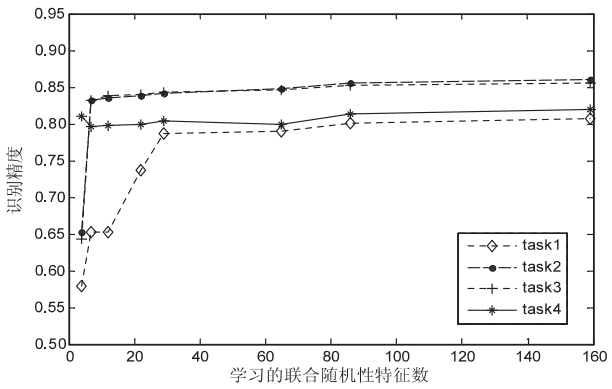

图 2 识别精度对比

4.2 识别性能分析

为了进一步验证该文提出的基于多任务特征学习的网络加密流量识别算法的有效性,与传统的单任务特征学习算法进行了对比。

传统的单任务特征学习算法不考虑任务之间的特征关联关系,通过求解 ℓ_1 范式正则化优化问题来进行单任务特征学习:

$$\min_W \sum_{i=1}^{n} \log(1 + \exp(-y_i(w^{\mathrm{T}}a_i + c))) + \rho \| W \|_1 \tag{19}$$

其中 $[a_1, a_2, \cdots, a_n] \in \mathbb{R}^{n \times d}$ 是随机性特征矩阵,y_i 是样本标签,W 是参数模型,c 是偏置,单任务特征学习通过 ℓ_1 正则化进行特征学习。单任务特征学习和多任务特征学习算法的平均识别精度对比如图 3 所示。由图 3 可以看出,多任务特征学习的平均识别精度超过 80%,单任务特征学习的平均识别精度明显低于多任务特征学习,但单任务特征学习的平均识别精度随着特征数变化的稳定性优于多任务特征学习。

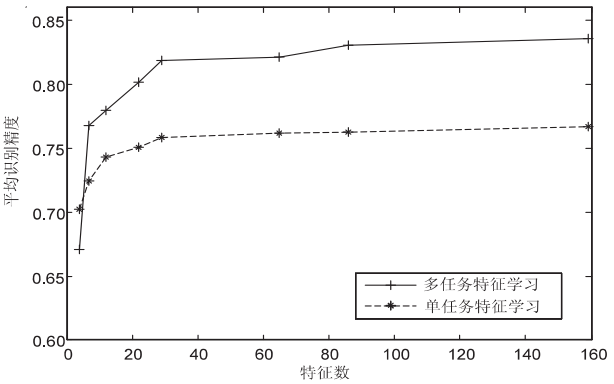

图 3 平均识别精度对比

5 结束语

基于多任务特征学习,将 NIST SP800-22 的 15 种随机性检验方法得到的检验数值作为 188 维随机性特征值,利用 $\ell_{2,1}$ 正则化项对一组相关任务进行联合特征学习,不仅能够准确识别已知加密协议流量,同时还对未知加密协议流量具有一定的识别能力。实验结果表明,提出的算法可以有效识别网络加密流量,平均识别精度超过 80%。

参考文献:

[1] MOORE A W, PAPAJIANNAKI K. Toward the accurate identification of network applications[J]. Lecture Notes in Computer Science, 2005, 3431: 41-54.

[2] KARAGIANNIS T, BROIDO A, FALOUTSOS M. Transport layer identification of P2P traffic[C]//Proceedings of the 4th ACM SIGCOMM conference on Internet measurement. Taormina, Siciy, Italy: ACM, 2004: 121-134.

[3] KANG H J, KIM M S, HONG J W. Streaming media and multimedia conferencing traffic analysis using payload examination [J]. ETRI Journal, 2004, 26(3): 203-217.

[4] MCGREGOR A, HALL M, LORIER P. Flow clustering using machine learning techniques[J]. Lecture Notes in Computer Science, 2004, 3015: 205-214.

[5] BACQUET C, GUMUS K, TIZER D, et al. A comparison of unsupervised learning techniques for encrypted traffic identification[J]. Journal of Information Assurance and Security, 2010 (5): 464-472.

[6] KUMANO Y, ATA S, NAKAMURA N. Towards real-time processing for application identification of encrypted traffic[C]// International conference on computing, networking and communications. Honolulu, HI: IEEE, 2014: 136-140.

[7] TREMBLAY M, KELOUWANI S, MELIN E. Method and system for identifying an application type of encrypted traffic: 8 539 221[P]. 2013-09-17.

[8] DALEY W M, SHAVERS C, KAMMER R. FIPS PUB. 140-2: security requirements for cryptographic modules[R]. [s. l.]: Booz-Allen and Hamilton Inc Mclean Va, 1999.

[9] RUKHIN A, SOTO J, NECHVATAL J. NIST special publication 800-22 revision 1a: a statistical test suite for random and pseudorandom number generators for cryptographic applications[EB/OL]. [2010-04-15]. http://csrc. nist. gov/groups/ST/toolkit/rng/documents/SP800-22rev1a. pdf.

[10] KILLMANN W, SCHINDLER W. A proposal for: Functionality classes and evaluation methodology for true (physical) random number generators[EB/OL]. [2001-09-25]. http://xepa15. fisica. ufmg. br/inetsec/uploadFiles/DOC/AIS_31_Functionality_classes_evaluation_methodology_for_true_RNG_e. pdf.

[11] 苏桂平,吕述望. 计算机安全系统中随机序列发生器的研究[J]. 计算机研究与发展,2003,40(7):994-1000.

[12] 范丽敏,冯登国,周永彬. 基于模糊评价的分组密码随机性评估模型[J]. 计算机研究与发展,2008,45(12):2095-2101.

[13] ARGYRIOU A, EVGENIOU T, PONTIL M. Multi-task feature learning[J]. Advances in Neural Information Processing Systems, 2007, 19: 41-48.

[14] ZHOU J, LIU J, NARAYAN V A, et al. Modeling disease progression via multi-task learning[J]. NeuroImage, 2013, 78: 233-248.

[15] GONG P, YE J, ZHANG C. Multi-stage multi-task feature learning[J]. Journal of Machine Learning Research, 2013, 14 (1): 2979-3010.

[16] GONG P, ZHOU J, FAN W. Efficient multi-task feature learning with calibration[C]//Proceedings of the 20th ACM SIGKDD international conference on knowledge discovery and data mining. [s. l.]: ACM, 2014: 761-770.

[17] GONG P, YE J, ZHANG C. Robust multi-task feature learning [C]//Proceedings of the 18th ACM SIGKDD international conference on knowledge discovery and data mining. Beijing, China: ACM, 2012: 895-903.

[18] ARGYRIOU A, EVGENIOU T, PONTIL M. Convex multi-task feature learning[J]. Machine Learning, 2008, 73(3): 243-272.

[19] CHEN J, ZHOU J, Ye J. Integrating low-rank and group-sparse structures for robust multi-task learning[C]//Proceedings of the 17th ACM SIGKDD international conference on knowledge discovery and data mining. San Diego, California, USA: ACM, 2011: 42-50.

[20] JI S, YE J. An accelerated gradient method for trace norm minimization[C]//Proceedings of the 26th annual international conference on machine learning. Montreal, Canada: ACM, 2009: 457-464.