基于相似性度量的网络流分类模型融合

姚永生,董育宁,邱晓晖
(南京邮电大学 通信与信息工程学院,江苏 南京 210003)

摘 要:由于网络流特征会随时间和网络环境的变化而发生概念漂移,不同类别应用的流发生漂移情况不同,导致基于机器学习的流量分类方法精度明显降低。同时,随着互联网网络技术的不断提高,使得过去采集并做好标签的大量视频流样本数据会发生很大变化,导致可用的训练集较少,需要实时采集和标注大量的新数据。针对上述问题,提出一种结合 Jensen-Shannon 距离、MultiTrAdaBoost 和 RandomForest 算法的分类方法。该方法的核心思想是:度量新老视频数据流之间的相似性,根据度量结果判断采用何种模型进行分类,其中的迁移学习分类方法是从老数据集中选出有用信息的样本来辅助新数据集样本的识别与分类。文中新老数据集样本特征属性分布是不一样的。实验结果表明,与现有的方法比较,该方法可以更好地实现典型的网络视频流分类,表现出较好的分类性能和泛化能力(即,模型的总体准确率标准差较小)。

关键词:Jensen-Shannon 距离;迁移学习;机器学习;网络流分类;概念漂移
中图分类号:TP301 **文献标识码**:A **文章编号**:1673-629X(2021)12-0007-06
doi:10.3969/j.issn.1673-629X.2021.12.002

Network Traffic Classification Model Fusion Based on Similarity Measurement

YAO Yong-sheng, DONG Yu-ning, QIU Xiao-hui
(School of Telecommunications & Information Engineering, Nanjing University of Posts and Telecommunications, Nanjing 210003, China)

Abstract: Because network flow characteristics will experience conceptual drift with time and network environment changes, the flow of different types of applications drifts differently, resulting in a significant reduction in the accuracy of the traffic classification method based on machine learning. Meanwhile, with the continuous improvement of Internet network technology, the large number of video stream sample data collected and labeled in the past will change greatly, resulting in fewer training sets available, and a large amount of new data needs to be collected and labeled in real time. Regarding the problem above, a classification method combining Jensen-Shannon distance, MultiTrAdaBoost and RandonForest algorithms is proposed. The core idea of this method is to measure the similarity between the new and old video data streams, and determine which model to use for classification based on the measurement results. The migration learning classification method is to select useful information samples from the old data set to assist the identification and classification of the new data set samples. In the article, the distribution of feature attributes of the new and old data sets is different. Experiment shows that compared with the existing methods, the proposed method can better implement typical network video stream classification, showing better classification performance and generalization ability (that is, the overall accuracy of the model has a smaller standard deviation).

Key words: Jensen-Shannon distance; transfer learning; machine learning; network traffic classification; conceptual drift

0 引 言

随着互联网技术的高速发展,网络视频流在网络通信中所占有的比例越来越高。种类不一样的网络应用对网络服务质量(quality of service, QoS)的要求不一样。视频流业务的细分类是实现端到端 QoS 的前提。互联网服务提供商为了给不同类型的视频应用分配合理的网络资源,需要将网络流应用进行细分类,更不能把所有网络业务当作同一个类型的类别区分。

近年来,机器学习(machine learning, ML)方法在网络数据流分类与识别中已得到广泛应用[1]。但遗憾的是经典的 ML 面对下面问题将无法很好地处理:首先,基于流特征的 ML 方法会随着复杂的网络环境发

收稿日期:2020-12-10 **修回日期**:2021-04-13
基金项目:国家自然科学基金(61271233)
作者简介:姚永生(1990-),男,硕士,CCF 会员(G7404G),研究方向为多媒体通信与视频业务流;董育宁,教授,博导,研究方向为多媒体通信与无线网络。

生变化导致网络流分布动态变化以及不同流特征间的差异引起网络流概念漂移;之前由老数据流量建立的 ML 模型,由于不同种类的网络流属性漂移情况不一致,使得 ML 模型的识别准确性下降。其次,互联网技术处在发展变化中,几年前采集的样本可能过时并无法满足现实需要,而再次采集并标注新样本会消耗不必要的成本。因此,该文提出一种基于相似性度量的分类模型来解决上述问题。该模型能够完成只有少量带标签新样本集与大量老样本训练集,以及数据流概念漂移不相同的情况下,更好地实现网络视频流分类。

本研究的基本思路是基于以下事实:通过采集的新老数据流对比发现,某些类别网络流在两个不同数据集上比较相似,而其他类别则有较大的差异。训练好的 ML 对老数据集可以很好地进行分类,而对于新数据集总体分类效果不理想。迁移学习(transfer learning,TL)可以利用有标注的源领域知识,来辅助目标领域的知识获取和学习。这样 TL 就能够利用之前学习到的知识,迁移到新数据集中完成差异较大类别的分类。鉴于此,该文的目的是研究何时采用 TL,何时采用 ML。根据 Jensen-Shannon(JS)距离[1]度量源域和目标域之间的相似性,距离越小,两者的相似性越大。通过相似性比较,选择采用何种模型。一方面,TL 可以解决数据集样本标注不足的问题,充分利用了源域的相关知识;另一方面,ML 模型是之前训练好的网络模型,对于相似性较大的后续流类别较适用,能够提高分类准确率和模型利用率。

特征选择方法在数据预处理阶段至关重要,通过选取出最优特征子集,能够有效降低后端分类算法的计算时间复杂度,同时可以更好地提高模型分类精度[2]。

该创新方法如下:

(1)结合 JS 距离、MultiTrAdaBoost[2] 和 Random Forest(RF)[1]算法提出了基于相似性度量的分类模型融合方法。该融合模型采用 JS 距离度量两个领域流分布的相似性,结合 ML 和 TL,可以较好地利用过去的数据,在节省标记新样本数据集成本的同时提高了模型的总体准确率。

(2)探究数据集中类别与模型选择之间的关系,不同种类的数据流特征发生漂移情况不一样。采集新的数据流进行特征提取,通过实验验证上述融合模型的有效性。

1 相关工作

1.1 网络视频流分类方法

Yang 等[3]根据不同类型的视频具有不同的下行传输速率变化模型模式,提出了一种基于 M 值概率分布并使用支持向量机的网络视频流分类算法。Garcia 等人[4]采用深度包检测分类流量,在流量到达率非常高的场景中,分类结果较好。Wu 等[5]针对种类不一致的网络流分布的差异性,提出一种可以完成存储低、延时小、准确率高的流细分类算法,同时实现了现实网络数据中较高分类识别率的效果。杨凌云等[6]应用短网络流包替代长网络数据包进行网络样本流分类,有效地降低了网络数据流算法时间复杂度并明显地提升了网络数据流分类准确率。

1.2 迁移学习分类方法

王彦等[2]提出了一种基于 SAMME[7] 和 TrAdaBoost[8]的 TL 分类算法,该算法能够对网络中不同种类的视频数据流进行有效地识别与分类,并有效节约了新数据集标记成本。Wang 等[9]阐述了类内迁移的思想,同时指出仅仅一个整体的特征迁移转换学习是不够的,当加入类内之间的相似性,类内特征能够用于类内迁移,从而实现更好的迁移学习。刘振等[10]提出了一种基于多重相似性的多源 TL 算法,能够从多个不同源域中挖掘更多的知识用于目标域学习,还可以根据域间的相似性有选择地进行迁移。当分布不同时,Cai 等人[11]运用 TL 来实现网络视频流量的两类分类,并改进了算法同时节约了计算时间成本。刘三民等[12]基于 TL 方法,应用过去老数据集有用知识来帮助目标域新模型知识学习,能够有效缓解新数据集标记样例的缺失。

2 预备知识

2.1 网络流概念漂移

在真实网络环境下,大量的数据流量高速率传输。随着复杂的网络环境发生变化导致网络流分布动态变化,进而导致采用网络流特征 ML 方法分类准确率会有不同程度地降低[13]。为此,网络流预测分类模型需要不断的迭代更新,变化的过程能够理解为网络数据流产生概念漂移的问题[14]。由概率理论,模型通过计算网络流的特征 $X=\{x_1,x_2,\cdots,x_n\}$ 识别其为某一类 Y 的概率来确定。流特征 X 是已知的,识别结果概率 $P(y/X)=\dfrac{P(y)P(X/y)}{P(X)}$。即期望函数 $f:X\rightarrow Y$:

$$H_g(x)=\underset{k}{\operatorname{argmax}}\sum_{t=1}^{N}(\ln(1/\beta_t))\prod(h_t(x)=k) \tag{1}$$

其中,分母 $P(X)$ 是样本统计特征 X 的概率,$P(X)=\prod_{i=0}^{n}P(x_i)$,$P(x_i)$ 表示训练集概率,网络流特征 x_i 将引起概率分布 $P(x_i/Y)$ 发生变化,进而引起 $P(Y/X)$ 变化;同时,先验概率 $P(Y)$ 也会引起 $P(Y/X)$ 变化。

2.2　JS 距离

JS 距离[1]可以用来描述源域和目标域之间的距离,进而衡量两个数据域之间的差异。该距离公式如下:

$$D_{JS}=\sqrt{H(\sum_{k=1}^{k=2}\frac{1}{2}M_k)-\sum_{k=1}^{k=2}H(M_k)} \tag{2}$$

其中,

$$H(M_k)=-\sum_{i=1}^{N}M_{k,i}\log M_{k,i} \tag{3}$$

N 为源域和目标域中特征向量的维数。

2.3　RF 算法

RF 算法是 Bagging 方法和决策树方法结合的集成 ML 算法。RF 通过随机建立多个决策树,每个决策树都是一个分类器,并将其合并到一起以投票的方式获得更准确和稳定的预测。

2.4　MultiTrAdaBoost 算法

MultiTrAdaBoost 算法[2]继承了 TrAdaBoost 的迁移思想,并结合 SAMME 来实现多分类,是一种通过训练多个弱基础学习器,最终生成强分类器的集成学习算法。该算法最终输出的函数为:

$$H_g(x)=\underset{k}{\operatorname{argmax}}\sum_{t=1}^{N}(\ln(1/\beta_t))\prod(h_t(x)=k) \tag{4}$$

其中,N 为循环的最大次数,h_t 为弱学习器预设函数,$\beta_t=\gamma_t/(K-1)$。$\prod$ 是一个指示函数,假设 $h_t(x)=k$,那么其值为 1,否则为 0。

2.5　MSGA 算法

MSGA[2]是一种基于 MultiSURF[15]和 GA(genetic algorithm)[16]的混合式特征选择算法。MultiSURF 是一种过滤式特征选择方法,可以实现特征尺寸的快速降维;GA 可以降低特征的冗余度,两者相结合可以选择出更优的特征子集。

2.6　GA 算法

GA 算法是一种全局性和自适应性的进化搜索算法,呈现了在自然选择过程中学习知识的不断主动获取和空间知识的不断自动搜索,能够加以对搜索过程的整体性进行有效地自适应调整并求得最佳解。在 ML 领域中,是一种过滤式的特征提取方法,去除冗余特征,得到最优特征子集。

3　文中方法

3.1　特征提取和选择

在网络流特征集提取过程中,首先依据 MultiSURF 算法计算出每个属性(例如:下行包大小均值)的权重值,按其权重进行排序,选择前 m 个属性。其次,用 m 个属性(特征)随机初始化原始种群(选取的特征数组成原始种群),并计算每一个个体的适应度函数值(选取分类学习算法 CART[2]的准确率用作适应度函数)。然后,利用 GA 算法,对个体进行选择、变异、交叉(选择适应度值较高的个体,交叉即是进行特征的互换,变异在一定程度上增加种群多样性),将新生成的后代加入到种群中,形成新的种群。当个体的适应度函数值不再变化,或算法达到最大的迭代次数,此时输出结果为最佳个体。在进行分类实验之前,使用 MSGA 特征选择算法从 25 个原始特征(见表 1)筛选出 8 个特征,如表 2 所示。

表 1　原始数据集 25 个特征

编号	特征名称	特征描述
1	downlink_up_rate_bytes	上下行字节速率之比
2	downlink_valid_ip_ratio	下行有效地址之比
3	downlink_up_counts_ratio	下上行数据包数目之比
4	downlink_up_sf_count_ratio	上下行片段数目之比
5	uplink_std_interval	上行包到达时间间隔标准差
6	uplink_subflow_features	上行包流片段特征
7	uplink_max_ps	上行包大小最大值
8	uplink_max_interval	上行包到达时间间隔最大值
9	uplink_interval_var	上行包到达时间间隔方差
10	uplink_ave_rate	上行包平均速率
11	datalink_avg_ps	整体包大小均值
12	datalink_valid_ip_ratio	整体包有效地址之比
13	datalink_std_ps	整体包大小标准差
14	datalink_std_interval	整体包到达时间间隔标准差
15	datalink_subflow_features	整体包流片段特征
16	datalink_interval_max	整体包到达时间间隔最大值
17	downlink_var_ps	下行包大小方差
18	downlink_ave_rate	下行包平均速率
19	downlink_max_ps	下行包大小最大值
20	downlink_ps_pdf_entropy	下行包大小信息熵
21	downlink_std_ interval	下行包到达时间间隔标准差
22	downlink_interval_var	下行包到达时间间隔方差
23	downlink_cv_rate	下行包速率均方值
24	downlink_interval_pdf_ent	下行包到达时间间隔信息熵
25	downlink_entropy_ps	下行数据包信息熵

表 2　MSGA 选出的 8 个特征

特征名称	特征描述
uplink_max_ps	上行包大小最大值
downlink_ up_rate_bytes	上下行字节速率之比
uplink_interval_var	上行包到达时间间隔方差
downlink_interval_pdf_ent	下行包到达时间间隔信息熵
uplink_std_interval	上行包到达时间间隔标准差
uplink_ave_rate	上行包平均速率
datalink_avg_ps	整体包大小均值
datalink_interval_max	整体包到达时间间隔最大值

3.2 混合式分类算法

首先利用 MSGA 算法筛选出 8 个特征,用老数据训练 ML 模型,再用老数据集加上一定比例的新数据训练 TL 模型。在训练 TL 模型过程中,不同比例的新样本数据的增加,经过每一轮的不断迭代学习,进而将老样本集中获取的有用知识应用到训练 TL 模型中,进一步提高融合模型分类精度。通过 JS 距离,衡量新老流特征之间的相似性,相似性越大,说明新老流之间的流属性发生概念漂移较小,采用已训练好的 ML 模型;反之新老流之间的相似性较小,采用 TL 模型,利用源域中学习的知识,进行分类预测。采用融合模型的分类准确率较单一 TL 模型有一定的提高。

提出的混合式分类方法 JSD-MTAB-RF 结合 JS 距离、MultiTrAdaBoost 和 RF 模型,其分类的计算过程如下:

(1)使用 MSGA 算法从原始数据集 25 个特征中筛选出 8 个特征。包含 8 个特征子集的数据集划分为老数据集 T_a(训练集),$a\%T_b(0 \leqslant a \leqslant 100)$ 的新数据集(训练集),$(1-a\%)T_b$ 的测试集(S)。

(2)用老数据集 T_a 训练 ML 模型,$T_a + a\%T_b$ 训练 TL 模型。

(3)比较新老流之间 JS 距离大小。

(4)当 JS-distance(S, $D(T_a)$)<JS-distance(S, $D(T_b)$)时,采用传统 ML(RF)模型;反之,采用 TL 模型。

(5)模型预测输出结果。

算法框图如图 1 所示。

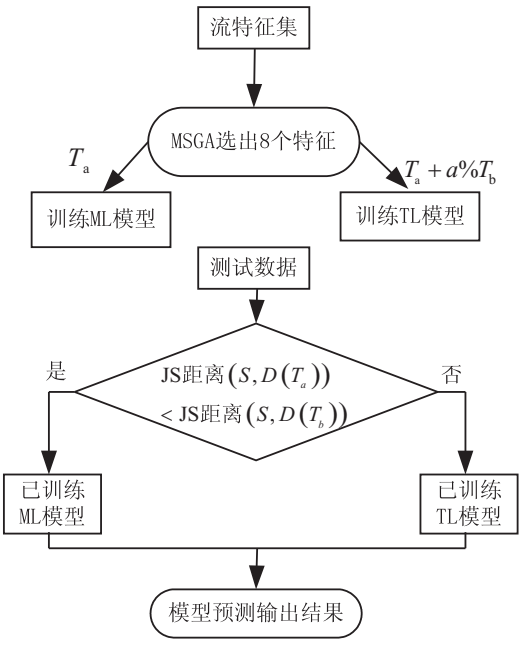

图 1 JSD-MTAB-RF 算法流程

具体实现算法的伪程序见算法 1。

算法 1:JSD-MTAB-RF。

Input:带标签训练样本 T_a 和 T_b 和测试集 S、应用类型数量 K、弱基础分类器、循环次数最大值 N 和旧模型 Rf(x)。

Output:类别预测结果。

1. 初始化权重向量, $W^1 = (w_i^1, \cdots, w_{n+m}^1)$

2. for $t = 1, 2, \cdots, N$ do

3. 归一化 $p^t = w^t / (\sum_{i=1}^{n+m} w_i^t)$

4. 使用弱基础分类器,输入组合训练样本 $T(T_a \cup T_b)$、权重分布 p^t 和测试集 S,获得在测试集 S 上的弱学习器 $h_t: X \rightarrow Y$。

5. 得出 h_t 在 T_b 上的错误率 ε_t:

$$\varepsilon_t = \sum_{i=n+1}^{n+m} w_i^t \prod (h_t(x_i) \neq c(x_i)) / \sum_{i=n+1}^{n+m} w_i^t$$

6. $\beta_t = \varepsilon_t / ((1-\varepsilon_t)(K-1))$

7. $$w_i^{t+1} = \begin{cases} w_i^t \beta^{\prod (h_t(x_i) \neq c(x_i))}, 1 \leqslant i \leqslant n \\ w_i^t \beta_t^{-1 \cdot \prod (h_t(x_i \neq c(x_i)))}, n+1 \leqslant i \leqslant n+m \end{cases}$$

8. 最终分类器:

$$H_g(x) = \operatorname{argmax}_k \sum_{t=1}^{N} (\ln(1/\beta_t)) \prod (h_t(x) = k)$$

9. for i in range(len(S) - 1)do

10. $$\operatorname{dis}(x, y) = \frac{1}{2} D(x \| M) + \frac{1}{2} D(y \| M)$$

11. 其中: $M = \frac{1}{2} * (x + y)$, $D(x \| M) = \sum_i x(i) \ln \frac{x(i)}{M(i)}$

12. if dis($S[i]$, Avg(T_a)) < dis($S[i]$, Avg(T_b))

13. output = Rf($S[i]$)

14. else output = $H_g(S[i])$

15. return output

16. end

4 实验结果及分析

4.1 数据集简介

在本实验中,由 Wireshark 抓包软件采集两个网络流数据集,实验中所抓取的网络视频流样本的时长均为 10 分钟。第一个数据集为收集 790 个样本的老数据集,于 2013 年 6 月在南京邮电大学校园网采集。第二个数据集为收集 458 个样本的新数据集,于 2019 年 9 月在南京邮电大学校园网采集。这两个数据集都包含 6 个视频应用类别:非对称式视频流:点播超高清视频(CD/1080p)、点播高清视频(HD/720p)与点播标清视频(SD/480p);以及对称式视频流:即时通信类视频(IVC)、网络类视频(P2P)和在线直播视频(ILV)。数据集中每一个样本均提取 25 个流特征(见表 1)和一个类标签。

对老数据集而言一些特征属性有用信息已过时,将其看作 T_a。新数据集可以划分成以下两块:一块用于模型训练的样本集 T_b,另一块用于模型的测试样本 S,同时两块样本集同分布。实验中,从新数据集中分别提取 20%,40% 和 60%(即 a = 20、40 和 60)的数据参与训练集,其余为测试集 S。

4.2　实验设置

硬软件平台是具有 Inter(R)Core(TM) i7-9750H、2.60 GHz CPU 和 16.0 GB RAM 的 PC 机,Win10 操作系统,运行 Python 编程语言(Python3.7 版本),在集成开发环境 Pycharm 中编写代码。

文中方法与文献[2]方法进行性能比较,后者将 T_a 和 T_b 组合形成训练集。实验采用 5 折交叉验证方法,测试指标给出均值和标准差。

4.3　分类评价指标

实验使用四个性能评估指标,分别为准确率(A)、查准率(P)、查全率(R)和 F1-测度(F1)。

(1)准确率:它是分类器准确分类的样本数与样本总数的比率。

$$A = \frac{TP + TN}{TP + TN + FN + FP} \tag{5}$$

(2)查准率:计算全部正确的项目(TP)与全部实际的项目(TP + FP)的比例。

$$P = \frac{TP}{TP + FP} \tag{6}$$

(3)查全率:计算全部正确的样本(TP)与应检索的全部样本(TP + FN)之比。

$$R = \frac{TP}{TP + FN} \tag{7}$$

(4) F1-测度:是查准率和查全率的加权平均值。

$$F1 = \frac{2 * P * R}{P + R} \tag{8}$$

4.4　实验结果

表 3 给出了文中方法与文献[2]方法的总体准确率比较。以 T_b 占比 20% 为例,文中方法的总体准确率为 95.4%,而文献[2]方法为 94.6%。究其原因,文中方法是利用 JS 距离度量新老流分布之间的相似性,根据距离的大小,判断相似程度,选择合适的学习模型;相比于文献[2]方法只使用 TL 模型,提高了总体性能。随着带标签的新数据集 T_b 在训练集中所占比例的提高,分类总体准确率有所提高。不过由于计算 JS 距离和模型的加载耗时,文中方法比文献[2]计算量略有增加。其中,训练时间增加0.13 s,识别时间增加 0.03 s。因训练阶段只运行一次,主要关注的是识别时间。

表 3　两种方法的总体准确率(均值±标准差)和运行(训练+识别)时间对比

方法	T_b 占比	总体准确率	训练+识别时间/s
文献[2]	20%	0.946±0.129 6	0.28+0.13
	40%	0.954±0.086 0	
	60%	0.977±0.124 9	
文中方法	20%	0.954±0.076 2	0.41+0.16
	40%	0.965±0.074 2	
	60%	0.983±0.072 8	

表 4 显示了文献[2]与文中融合模型 JSD-MTAB-RF 方法在种类不一样的网络数据流中的预测结果。在新老样本数据中,数据样本差异性大小是影响性能的根本原因。随着时间的变化,新老数据集中流之间的属性会发生概念漂移;不同种类的流漂移的情况不一样,某些流特性漂移波动较大,而某些流特性漂移波动较小;TL 对于相似性较小的流类别适用,对于相似性较大的流类别不适用。在 ILV 类别,文中方法的 F1 测度、查准率、查全率都有所提高,是因为该类新老数据相似性较大,采用了训练好的 ML 模型,较只采用 TL 模型性能有所提升。同时,随着带标签的新数据集 T_b 在模型训练中占比越高(a = 40、60)时,文中方法对部分视频类型分类评价指标有所提升。

表 4　两种方法对网络视频流的分类性能(均值±标准差)对比(T_b 占比 20%、40%、60%)

视频类型	方法	查准率	查全率	F1
SD	文献[2]	1.0±0.077	0.93±0.032	0.96±0.032
	文中方法	1.0±0.045	0.93±0.032	0.96±0.032
		1.0±0.045	0.93±0.032	0.96±0.032
		1.0±0.045	0.95±0.033	0.98±0.034
HD	文献[2]	0.85±0.032	0.90±0.055	0.87±0.032
	文中方法	0.88±0.032	0.90±0.045	0.89±0.032
		0.88±0.032	0.90±0.045	0.89±0.032
		0.92±0.033	0.90±0.045	0.93±0.035
CD	文献[2]	0.85±0.000	1.0±0.032	0.92±0.000
	文中方法	0.85±0.055	1.0±0.000	0.92±0.000
		0.86±0.056	1.0±0.000	0.93±0.012
		0.88±0.056	1.0±0.000	0.95±0.016
IVC	文献[2]	1.0±0.000	1.0±0.032	1.0±0.000
	文中方法	1.0±0.000	1.0±0.000	1.0±0.000
		1.0±0.000	1.0±0.000	1.0±0.000
		1.0±0.000	1.0±0.000	1.0±0.000
P2P	文献[2]	1.0±0.000	1.0±0.032	1.0±0.000
	文中方法	1.0±0.000	1.0±0.000	1.0±0.000
		1.0±0.000	1.0±0.000	1.0±0.000
		1.0±0.000	1.0±0.000	1.0±0.000
ILV	文献[2]	1.0±0.032	0.86±0.032	0.92±0.032
	文中方法	1.0±0.000	0.89±0.071	0.93±0.045
		1.0±0.000	0.92±0.073	0.95±0.047
		1.0±0.000	0.95±0.075	0.97±0.049

4.5　实验探究与验证

查看实验数据集中,六种类别的样本分别采用了何种学习模型。统计结果得出,IVL 类别样本全部采用了 ML 模型;因为,该类别的流特性概念漂移较小,新老流之间的相似性较大;而其他五个类别的样本几

乎全部采用了 TL 模型,因为这些类别的流特性概念漂移较大,新老流之间的相似性较小。

为了验证上述结论的正确性,作者于 2020 年 9 月采集了 100 条 ILV 类别的样本,利用本实验组成员编写的 NetFlowAnaLab 平台从 5 元组数据流中提取出 25 个原始特征和标注类别标签。之后用 MSGA 算法提取出 8 个特征(见表 2)用于实验。结果得出,对于新采集的 100 条 ILV 类别样本,采用 ML 模型分类准确率为 100%,而采用 TL 模型仅为 97.6%。

5 结束语

该文提出了一种结合 JS 距离、MultiTrAdaBoost 和 RF 的混合式分类方法 JSD-MTAB-RF,能够实现多种不同网络应用视频流识别与分类。实验结果表明,在新数据集和老数据集特征分布不完全相同的情况下,即网络流特征属性发生了概念漂移,提出的基于相似性度量的混合模型,相对于现有方法,尽管耗时略有增加,但准确率更高。未来依然有一些地方需进一步实验探讨。下一步的研究工作在具有不同特征空间分布的 TL 方法,可以通过特征变换,将两域之间的数据特征有效地转换到相同标准的特征空间中去,从而实现基于特征的 TL 方法。

参考文献:

[1] GARCIA J, KORHONEN T. Efficient distribution-derived features for high-speed encrypted flow classification[C]// Proceedings of the 2018 workshop on network meets AI & ML. New York, NY: ACM, 2018: 21-27.

[2] 王 彦,董育宁,葛 军. 实现网络视频流多分类的迁移学习算法[J/OL]. 计算机工程与应用,2020,56(10):1-6.

[3] YANG L Y, DONG Y N, WANG Z J, et al. Network video traffic classification based on probability distribution of m value[J]. Journal of Electronics and Information Technology, 2018, 40(5): 1094-1100.

[4] GARCIA J, KORHONEN T, ANDERSSON R, et al. Towards video flow classification at a million encrypted flows per second[C]//2018 IEEE 32nd international conference on advanced information networking and applications (AINA). Krakow, Poland: IEEE, 2018: 358-365.

[5] WU Z, DONG Y N, YANG L Y, et al. A new structure for internet video traffic classification using machine learning [C]//2018 sixth international conference on advanced cloud and big data (CBD). Lanzhou, China: IEEE, 2018: 220-224.

[6] 杨凌云,冯友宏,闫 鹤. 在线长视频流的短数据包分类[J]. 电声技术,2020,44(2):30-33.

[7] CHEN G, ZHANG Y, TANG D. A noise classification algorithm based on SAMME and BP neural network[C]//2018 IEEE 3rd international conference on big data analysis (ICBDA). Shanghai, China: IEEE, 2018: 274-278.

[8] HUANG X, RAO Y, XIE H, et al. Cross-domain sentiment classification via topic-related TrAdaBoost[C]//National conference on artificial intelligence. [s. l.]: [s. n.], 2017: 4939-4940.

[9] WANG J, CHEN Y, HU L, et al. Stratified transfer learning for cross-domain activity recognition[C]//IEEE international conference on pervasive and communication. Athens, Greece: IEEE, 2018: 1-10.

[10] 刘 振,杨俊安,刘 辉,等. 基于域相关性与流形约束的多源域迁移学习分类算法[J]. 计算机应用研究,2017,34(2):351-356.

[11] CAI L, JING X, SUN S, et al. P2P traffic identification based on transfer learning[C]//2013 IEEE international conference on granular computing (GrC). Hangzhou, China: IEEE, 2013: 22-26.

[12] 刘三民,刘余霞. 基于实例迁移的数据流分类挖掘方法[J]. 信息与控制,2019,48(3):380-384.

[13] JIN Y, DUFFIELD N, ERMAN J, et al. A modular machine learning system for flow-level traffic classification in large networks[J]. ACM Transactions on Knowledge Discovery from Data, 2012, 6(1): 1-34.

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

[15] URBANOWICZ R J, OLSON R S, SCHMITT P, et al. Benchmarking relief-based feature selection methods for bioinformatics data mining[J]. Journal of Biomedical Informatics, 2018, 85: 168-188.

[16] 戴晓晖,李敏强,寇纪淞. 遗传算法理论研究综述[J]. 控制与决策,2000,15(3):263-268.