·体系结构与软件技术· 文章编号：1000-3428(2019)04-0018-07 文献标志码：A 中图分类号：TP391

SDN 架构下的空间信息网络业务识别技术

潘成胜[1]，刘 勇[1]，石怀峰[1,2]，杨 力[1]

（1. 大连大学 信息工程学院 通信与网络重点实验室，辽宁 大连 116622；2. 南京理工大学 自动化学院，南京 210094）

摘 要： 针对传统的离线业务流量识别方法消耗时间长、实时性差的问题，通过对空间信息网络管控和网络资源的高效编排，提出一种基于软件定义网络（SDN）架构的空间信息网络业务识别技术。运用 OpenFlow 协议在线收集业务流量，提取流中前 5 个数据包作为一条子流，在 SDN 控制器上实现基于机器学习的在线业务分类，同时给出一种具有噪声过滤功能的协同训练算法 Dif-TriTraining。实验结果表明，与传统的 Tri-Training 算法相比，该算法能够有效提升业务识别的准确率。

关键词： 软件定义；空间信息网络；业务识别；噪声过滤；协同训练

中文引用格式： 潘成胜，刘勇，石怀峰，等. SDN 架构下的空间信息网络业务识别技术[J]. 计算机工程，2019，45(4)：18-24.

英文引用格式： PAN Chengsheng，LIU Yong，SHI Huaifeng，et al. Spatial information network business identification technology under SDN architecture[J]. Computer Engineering，2019，45(4)：18-24.

Spatial Information Network Business Identification Technology Under SDN Architecture

PAN Chengsheng[1]，LIU Yong[1]，SHI Huaifeng[1,2]，YANG Li[1]

（1. Key Laboratory of Communication and Network，School of Information Engineering，Dalian University，Dalian，Liaoning 116622，China；2. School of Automation，Nanjing University of Science and Technology，Nanjing 210094，China）

【Abstract】 Aiming at the problem that the traditional offline business traffic identification method consumes a long time，and has poor real-time performance，through the management and control of spatial information network and the efficient arrangement of network resources，a business identification technology of spatial information network based on Software Defined Network（SDN）architecture is proposed. The OpenFlow protocol is used to collect business traffic online，and extract the first five data packets in the flow as a sub-flow，and implement online business classification based on machine learning on the SDN controller. At the same time，a collaborative training algorithm Dif-TriTraining with noise filtering function is presented. Experimental results show that compared with the traditional Tri-Training algorithm，the algorithm can effectively improve the accuracy of business identification.

【Key words】 software definition；spatial information network；business identification；noise filtering；collarborative training

DOI： 10.19678/j.issn.1000-3428.0049794

0 概述

空间信息网络具有业务种类繁多、通信环境复杂、星上资源受限以及拓扑动态变化等特性，这对网络的灵活管控和业务的高效编排提出挑战。因此，建立业务流量特征到业务类型的映射关系，完成空间信息网络业务识别，对流量工程、网络管理以及业务编排等具有十分重要的意义。现有的业务流量识别方法多数在离线状态下进行，通过从网络中采集数据并存储，之后进行特征统计，并按照分类方法进行分类实验。在空间信息网络中存在大量的实时性业务，迫切需要能够以实时或近实时方式识别出不同的业务类型，而在线流量分类技术[1-2]能够快速识别出网络中流量类型，对于实现空间信息网络业务高效编排具有十分重要的意义。

现有的业务识别[3]技术主要有基于端口识别技术、深度包检测技术和机器学习等。随着动态端口技术、端口伪装技术以及无注册端口号应用程序的出现，传统基于端口识别方法的分类精度逐渐降低，深度包检测技术需要对数据包进行解析，但对于私

基金项目： 装备预研领域基金（6140449XX61001）。

作者简介： 潘成胜（1962—），男，教授、博士生导师，主研方向为软件定义网络架构、空地一体化网络体系结构、网络协议；刘 勇，硕士研究生；石怀峰，讲师；杨 力，教授。

收稿日期： 2017-12-21 **修回日期：** 2018-02-05 **E-mail：** 674019843@qq.com

密性要求较高的空间信息网络并不适用。机器学习是当前比较热门的技术之一,基于机器学习的业务分类技术在保证一定识别精度的基础上,可以充分保护用户隐私,成为当前业务识别技术研究的热点。

1998 年,BLUM A 和 MITCHELL T 提出基于机器学习的协同训练算法 Co-Training,该算法在 2 个充分冗余视图上分别建立 2 个分类器,使用分类器迭代,每次迭代将置信度高的样本进行标记,然后放至另一分类器的训练集中用于强化学习。然而,在多数情况下,该算法难以找到充分冗余视图的样本集,导致该方法无法适用于所有的训练样本。2007 年,南京大学学者周志华和蔡铭提出了 Tri-Training 算法,从已标记样本集中随机抽取 3 个样本子集,采用相同的分类算法训练出 3 个基础分类器,选取出一个作为主分类器,剩余的作为辅助分类器,取出 2 个辅助分类器分类结果相同的样本进行标记,然后用于辅助主分类器的强化训练,该算法不需要充分冗余视图,分类精度有所提升。

本文针对传统协同训练算法存在的基础分类器分类精度低、容易引入噪声数据等问题,提出一种改进 Dif-TriTraining 算法,以在噪声过滤功能基础上提升分类器的分类效果。

1　软件定义空间信息网络

软件定义网络(Software Defined Network,SDN)通过实现控制和数据转发的分离,增强网络的可重用性,降低网络设备的复杂度,提高网络的灵活性,更加有利于实现对网络的管理和控制。国内外开展了在空间信息网络中运用 SDN 技术的相关工作,文献[4]提出一种基于 OpenFlow 的软件定义卫星网络(Soft Defined Satellite Network,SDSN)架构设计方案,使用 GEO 天基系统作为分布式卫星系统(DSS)的控制层面,实现和数据层面的分离,提高了 DSS 网络的可重用性、高效性以及灵活性。文献[5]提出一种基于 SDN 的空间信息网络多控制器部署策略 MDD-SDSIN,该策略充分考虑了空间信息网络拓扑周期性变化和流量的突发性,并针对过载导致的控制器失效以及欠载导致的资源浪费问题,提出一种基于双门限的交换机动态迁移算法。实验结果表明,该算法能够提升系统的吞吐量,降低了系统的响应时延,提高网络的整体性能,保证了控制器间的负载均衡问题。文献[6]提出一种基于 SDN 架构的海军通信网络,解决当前海军网络中存在的多个卫星通信链路的共享和负载均衡以及带宽限制等问题,并给出一种多路径传输控制协议用以改善链路的中断问题。文献[7]提出可在机载 SDN 平台和地面网关上实现的软件,以确保该架构实现业务的可靠性、交付保证和时间要求。SDN 通过内置的信息收集机制和网络的灵活可编程性为业务分类领域开辟了新思路。文献[8]提出一种部署在企业网络中的架构,使用 OpenFlow 协议收集流量数据,获取数据集,并展示了如何将机器学习应用于流量分类,验证了有监督机器学习对数据集的高精度分类。本文使用 SDN 控制器收集来自 OpenFlow 交换机转发过来的流量,然后使用机器学习的方法进行流量分类。由于 GEO 卫星相对地面接收站的静止使得星地链路更加稳定,且 GEO 的覆盖率更高,因而在该 SDN 架构中采用 GEO 卫星作为控制层,MEO/LEO 层卫星履行交换机职责,SDN 架构的业务感知技术如图 1所示。

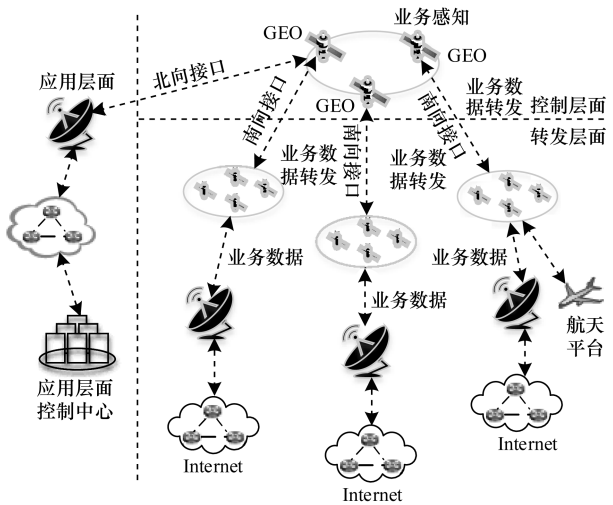

图 1　基于 SDN 架构的业务感知技术

该架构主要功能如下:

1)最左侧为应用层面,该层面可以部署在地面也可以部署在某一个固定空间站中。在该层面可以进行编程等操作,通过北向接口将定义的新软件信息在控制层面进行更新。

2)控制层面由 GEO 卫星组成,通过 OpenFlow 协议来控制数据层中子系统的接入方式及路由信息等,在控制层面收集数据层转发过来的数据包,然后再进行流量分类处理,该层是进行业务感知的关键。

3)数据层面包含 MEO/LEO 层和一些地面基础设施,在 MEO 和 LEO 卫星上部署 OpenFlow 交换机,并可以与地面接收站和控制层进行数据交互,如有业务经过交换机,交换机会将业务数据转发到控制器上,该层是实现业务感知的基础。在实现交换机转发数据包至控制器之前,需要通过基于 OpenFlow 协议的 Hello 消息来建立控制器与交换机之间的连接关系,使用 Echo 消息来确认交换机与控制器之间的连接状态。

2　空间信息网络业务识别技术

2.1　基于流中前 5 个包的流量识别技术

传统的业务识别方式是采用离线采集流量数据,分析相关流量特征进而展开相关研究,但该方式并不适用于实时性要求较高的空间信息网络。

流的初始阶段是建立通信协商的过程,这些数据包通常是一个预定义的消息序列,该消息序列在不同的应用程序之间是不同的,对此,文献[9]提出一种只针对 TCP 流的前几个数据包的在线流量分类方法,该分类方法分为离线训练和在线分类 2 个阶段,使用 K-means 聚类算法进行实验,结果显示,几乎所有应用均达到了 80% 以上的识别率。研究发现,基于前 5 个数据包就能实现簇间应用的最佳分离,50 个簇是实现应用分离和复杂度的最佳折中。文献[10]针对流中前 3 个数据包的相关特征进行分析,这些特征主要包括数据包大小、相邻 2 个数据包到达时间间隔、往返时延等特征。研究显示,流量的早期特征载有较为丰富的应用行为特征信息,可用于实现网络流量的识别研究,其中 TCP 流中的第 1 个数据包大小对于流量识别具有十分重要的意义。文献[11]利用流开始的前 5 个数据包统计相关特征,通过分析 3 种机器学习算法(C4.5、BayesNet 和 NBTree)分类结果,研究可用于在线流量分类的特征以及这些特征应该满足的条件,实验结果表明,3 种机器学习均取得较高的分类准确率(92% 以上)。

基于上述研究,本文选取流中前 5 个数据包作为一条子流进行相关特征的统计,以用于在线业务流量识别研究。

2.2 业务流量在线识别

传统离线流量分类方法必须等整条流都结束后才能进行分类处理,无法满足一些实时性要求较高或需要高速处理的业务需求。因此,本文设计一种在线学习框架,如图 2 所示。

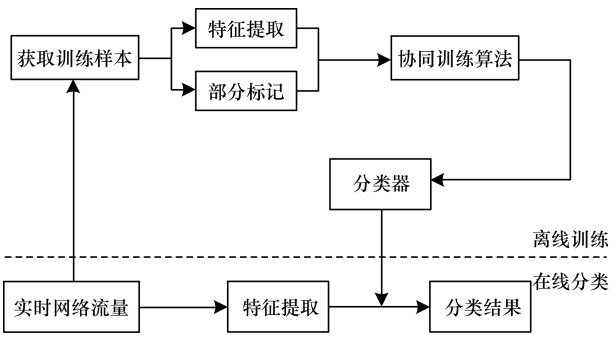

图 2 在线流量分类流程

在线识别分为离线训练和在线分类 2 个阶段。

1)离线训练阶段

离线训练阶段主要获取实时业务流量数据作为训练集,通过提取流量特征和对样本进行部分标记,实现基于 Dif-TriTraining 算法分类器训练。详细步骤如下:

(1)在线抓取实时流量数据获取训练样本;

(2)对训练样本进行特征提取并按比例对其中部分样本进行标记,然后放入有标记样本集中,剩余的未标记样本放入未标记样本集中;

(3)使用 Dif-TriTraining 算法来进行训练获得分类器。

2)在线分类阶段

该阶段的主要工作在于抓取实时网络业务流量进行特征提取,使用分类器进行流量分类,以获得分类结果。详细步骤如下:

(1)在线获取实时业务流量数据作为测试样本;

(2)对测试样本进行特征提取;

(3)使用分类器进行分类实验获得分类结果。

2.3 基于在线业务流量识别的特征选择

2.3.1 在线流量分类对流量特征的要求

传统的基于离线的流量分类特征[12]并不完全适用于在线流量分类,需要在空间信息网络链路上,在尽可能短的时间内识别出业务的类型。用于在线业务流量分类的特征应遵循以下条件:

1)尽早分类

在线业务流量分类要求能够尽快地识别出流量类别,这就需要从流开始的若干个尽可能少的数据包中统计特征,因为在空间信息网络中存在着大量实时性要求较高的业务,如遥控、遥测、语音视频等业务,如果不能尽早识别出这些流量,那么流量分类的实用性和意义就大打折扣。因而,诸如流的持续时间、流长度这类必须要求完整地观测才能得出的特征,不适用于在线分类。

2)计算存储开销低

在线业务流量分类要求特征必须具有较低的计算开销和存储开销,以保证具有最小的处理时延和尽可能大的吞吐量。由于空间信息网络带宽存储等资源受限,一些通过大量复杂计算而得到的统计特征将消耗大量的计算和存储资源,因而不适用于空间信息网络。

3)可重新训练分类器

在线业务流量分类要求特征必须有利于快速地重新训练分类器。网络流量动态特性的存在要求分类器能够及时更新自己的分类特征库,随着空间信息网络的快速发展,网络业务更新快,新类型业务不断出现,也需要通过重新训练分类器来识别新类型的业务。

2.3.2 在线流量分类特征选取

在在线网络流量分类中,统计特征的选取是其中一个至关重要的环节,其会直接影响分类的效果和性能。网络流量特征主要分为基于包和基于流[13]的分类特征。

1)基于包的统计特征:常用的基于包的流量特征包含包大小、相邻两数据包到达时间间隔和速率等。文献[14]选取网络流中前 n 个数据包的大小和方向作为流特征进行流量识别研究,分类结果达到 80% 以上。文献[15]选取 TCP 流中前 3 个包大小作为分类特征,并将服务器端口号作为辅助特征,结果表明,对一些特征进行组合能够达到 97% 的分类准确性。文献[16]使用 249 条流特征来标记一条 TCP 流构建了数据集,该数据集为当前比较权威的

数据集之一,然而在该数据中存在着大量通过傅里叶变换等复杂计算得到的特征值,这些特征并不适用于空间信息网络。空间信息网络中的高时延特性可能会导致包到达顺序不一致的情况,因而诸如流中某一位数据包大小的统计特征也不适用。基于此,本文选取子流中最大包、最小包、平均包的大小、平均到达时间以及到达时间间隔均值来作为统计特征,以用于空间信息网络在线业务识别研究。

2)基于流的统计特征:常用的基于流的统计特征包括流的大小、流的持续时间、标志位个数等。文献[17]分别利用网络中前 5 个和前 25 个数据包的 10 项特征对流量进行分类,结果表明,后者的分类准确性能达到 95% 左右。文献[18]从 38 个流特征中选取最佳的 4 个特征通过聚类算法获得约 94% 的分类准确率。人们也常使用流中 ACK 包数和一些特殊标志位作为统计特征,由于空间信息网络采用了不同的传输协议,导致数据包格式有所不同,本文不再使用这些统计特征,只使用子流的持续时间和子流大小作为统计特征。

基于以上分析,本文选取了易于统计、计算复杂度较低且适用于空间信息网络在线流量分类的统计特征(见表 1),这些特征是通过对子流(前 5 个数据包)的相关信息进行统计而得出。

表 1　适用于在线业务识别的特征

序号	基本特征	描述
1	流大小	子流的总数据包大小
2	平均包大小	平均包大小
3	最大包	最大包大小
4	最小包	最小包大小
5	间隔时间	平均达到时间
6	流持续时间	子流的持续时间
7	最小时间间隔	相邻两数据包到达时间间隔最小值

3　Dif-TriTraining 算法

在真实的空间信息网络环境中,网络流量存在大量的噪声,其噪声主要包含:

1)由网络环境所引入的噪声,如 Ka 频段受雨水、天气、太阳活动等而引起的高误码率,还有空间信息网络的长时延等问题而引入大量的噪声,这些噪声会对训练分类器造成影响,从而降低分类器的分类精度。

2)传统的 Tri-Training 协同训练算法[19]本身还存在一些不足之处,在训练辅助分类器时可能会错误标记,从而引入标记噪声,使用被标记错误的样本来训练分类器也会对分类器的精度造成影响。

基于噪声学习理论,文献[20]提出,如果大部分被标记的样本被标记正确,那么引入的错误标记样本所带来的分类错误率会被抵消。对此,本文提出一种样本差异性度量方式计算样本与整个样本集之间的差距,用以衡量样本在整个样本集中的偏移情况,通过计算值大小来确定样本标记置信度。此外,还引入了抽取比例的概念,用于选取大量标记置信度较高的样本以抵消错误标记率所导致的分类错误率。基于此,本文提出了一种具有噪声过滤功能的 Dif-TriTraining 协同训练算法,来获取 2 个辅助分类器判决一致的样本,然后计算样本的标记置信度,若标记置信度越高则其被标记正确的概率就较大,从而可以选取出大量正确标记的样本用于训练分类器,避免了 2 个辅助分类器判决一致即认为正确的情形,有效降低由标记错误所带来的分类错误率。

在机器学习中为估算不同样本之间的差距,需要通过计算样本间距离的方式来实现,采用的方法主要有欧式距离、曼哈顿距离、切比雪夫距离等。本文采用了基于欧式距离的样本距离计算方式,假设有 2 个 n 维向量 $\boldsymbol{s}_1(x_{11},x_{12},\cdots,x_{1n})$ 和 $\boldsymbol{s}_2(x_{21},x_{22},\cdots,x_{2n})$,它们之间的距离为:

$$d_{12}=\sqrt{\sum_{k=1}^{n}(x_{1k}-x_{2k})^2} \tag{1}$$

通过欧式距离可以反映出 2 个样本之间的距离,但无法代表某一样本与整个样本集之间的整体性差距,需要计算样本与样本集之间的距离来筛选可靠性高的样本,本文定义了一种衡量单个样本与整个样本之间整体性差距的计算方式。

定义　通过单个样本与其他样本之间距离之和的方式来衡量单个样本与整个样本集之间的距离,将其称为样本差异性度量。

由此,需要计算单个样本与样本集中其他所有的样本之间的距离然后求和,即可得出该样本与整个样本集的距离,其计算公式如下:

$$d_i=d_{i1}+d_{i2}+\cdots+d_{iM}=\sum_{j=1}^{M}d_{ij}=\sum_{j=1}^{M}\sqrt{\sum_{k=1}^{N}(x_{ik}-x_{jk})^2} \tag{2}$$

其中,d_i 表示样本集中第 i 个样本与整个样本集之间的差距,M 代表样本总数,N 代表特征维度。

根据以上分析,本文将样本差异性度量思想应用到算法中,算法的主要流程如图 3 所示。

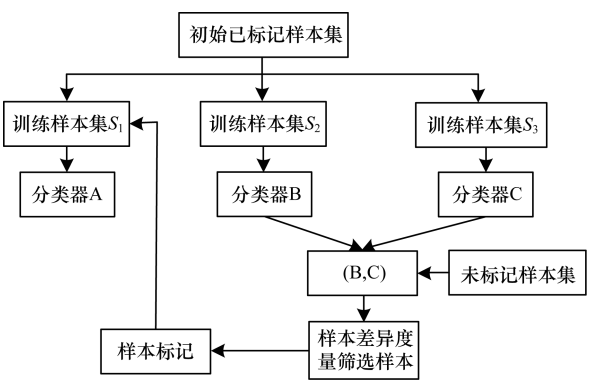

图 3　Dif-TriTraining 算法流程

假设初始已标记样本集为 L,未标记样本集为 U,训练分类器的具体步骤如下:

1)对有标记样本集 L 采用传统的可放回随机抽样方法(Bagging),从初始已标记样本集中获取 3 个有差异性的已标记样本集,再使用 C4.5[21] 分类算法对这 3 个有标记数据集进行训练,得到 3 个初始分类器 A、B、C。

2)选择其中任意一个为主分类器,假设 A 为主分类器,B 和 C 为辅助分类,然后利用辅助分类器 B 和 C 对未标记样本集 U 进行分类,做分类结果标记,将标记相同的样本与相应的标记组合成集合 X_a。统计集合 X_a 样本个数,假设为 K_1。

3)计算每个样本 x_i 属于 X_a 的样本差距。样本差距计算公式如下:

$$dis(x_i)=\sum_{j=1}^{M}\sqrt{\sum_{k=1}^{N}(x_{ik}-x_{jk})^2} \tag{3}$$

其中,$dis(x_i)$ 表示样本 x_i 的样本差距,N 表示样本的维度,x_{ik} 表示 x_i 第 k 维上的值。从样本差距的定义可知,$dis(x_i)$ 越小,意味着 x_i 与其他样本的相似度就越高,其标记置信度也就越高。因此,$dis(x_i)$ 越小,x_i 样本被正确标记的可能性就越大。

4)从 X_a 中选择 K_2 个 $dis(x_i)$ 值最小的样本,并将辅助分类器对该样本的标记作为其标记,将标记后的样本加入主分类器对应的已标记样本集进行扩展,然后重新训练该分类器,获得 X'_a。K_2 的计算公式为:

$$K_2=[select\_rate\times K_1] \tag{4}$$

其中,$select\_rate\in(0,1)$ 是样本差异较小的样本所占比例。

5)训练结束,X'_a 则为训练后最终的分类器。

4 实验结果与分析

4.1 实验数据集及实验环境

为验证本文所提出算法的有效性,采用新西兰奥克兰大学采集的业务流量数据集,记作 Auckland_Set。该数据集通过采集奥克兰大学园区边界路由器而获得。选取其中标号为 20010221-020000-0 和 20010221-020000-1 的 2 个流量数据集,包含了 2001-02-21 02:00:00 开始的 24 h 的流量信息,具体信息如表 2 所示。统计的应用类别如表 3 所示。

表 2 流量数据统计

数据集	数据包数量	数据大小/MB
2001010221-020000-0	3.1×10^7	920
2001010221-020000-1	3.3×10^7	986

表 3 数据集中的应用类别

应用类别	应用名称
WWW	Web browsers
BULK	FTP
CHAT	IRC,MSC Messenger
P2P	BitTorrent,eDonkey,Gnutella
INTERACTIVE	SSH,klogin,telnet
MAIL	IMAP,POP2/3,SMTP
MULIMEDIA	Windows Media Player
GAMES	Half-life,Counter-Strike
VOIPS	Asterisk
NEWS	NNTP
ATTACK	Internet worm
OTHER	—

为减少实际的工作量,每个应用(除了 OTHER 类)抽取 2 000 条,去除相关负载信息后作为本文实验的仿真数据。实验中采用的工具是 Matlab_R2012a,Matlab 函数库中已增加了对本文实验所需的决策树相关 API 的支持,在此基础上实现自己的机器学习方法。实验平台是实验室的 DELL 台式机,CPU 为 Intel(R) Core(TM) i7-3770 CPU @ 3.04 GHz,8 GB 内存,运行 Windows 7 操作系统。

为验证本文所提出算法的去噪性能,需要对原始的未标记训练样本进行加噪处理,统计每种应用的特征值的范围,随机选取 15% 的数据,根据每种应用类型对应的特征值范围进行随机性改变,作为加噪数据用于实验分析。

4.2 实验评价指标

为衡量本文提出算法的性能,采用以下 3 个指标来进行评价,分别为整体识别率 *Total_Accuracy*、精度 *Precision* 和召回率 *Recall*,其定义如下:

$$Total\_Accuracy=\frac{\sum_{i=1}^{m}TP_i}{\sum_{i=1}^{m}TP_i+FN_i} \tag{5}$$

$$Precision=\frac{TP_i}{TP_i+FP_i} \tag{6}$$

$$Recall=\frac{TP_i}{TP_i+FN_i} \tag{7}$$

其中,TP_i 代表实际类型为 i 被判定为类型 i 的样本数,TN_i 代表实际类型为非 i 的样本被判定为非 i 类型的样本数,FP_i 代表实际类型为非 i 的样本被判定为类型 i 的样本数,FN_i 代表实际为 i 类型的样本被判定为非 i 类型的样本数。

整体识别率用于描述被正确识别出的类型的样本数与整个样本之间的比值,强调的是整体的结果,不涉及具体的应用类型,而精度和召回率则涉及具体类型的相关计算。

4.3 结果分析

本文从所获得的数据中随机抽取 20% 作为测试

集,剩余的 80% 作为训练集。然后在此基础上进行实验分析。

实验 1　从训练集中抽取 30% 的样本作为有标记样本,剩余 70% 的样本作为辅助训练样本,取 *select_rate* 的值为 0.5,其中辅助训练样本分为两类:一类是原始数据样本;另一类是加噪数据样本。采用传统的 Tri-Training 算法分别运用 2 个辅助训练样本来训练分类器,然后使用测试集来验证加噪声后对训练分类器的影响,结果如表 4 所示。

表 4　加噪前后的识别结果每列平均分布　%

数据	整体识别率	精度	召回率
原始数据	90.7	90.9	90.8
加噪数据	84.8	84.9	84.8

从表 4 可以看出,加入噪声数据后算法的整体识别率降低 6%,这说明在大量噪声存在的情况下会明显影响算法的识别效果。

实验 2　保持实验 1 中所使用的参数不变,只使用加噪数据,分别使用 Dif-TriTraining 算法和 Tri-Training 算法来训练分类器,然后根据分类结果进行分析,验证本文算法的去噪性能,结果如表 5 所示。

表 5　不同识别算法的结果分析　%

算法	整体识别率	精度	召回率
Dif-TriTraining 算法	92.5	92.4	92.6
Tri-Training 算法	84.9	85.0	84.9

表 5 分别运用 2 种算法对加噪过后的数据进行分类实验,使用 Dif-TriTraining 算法在整体识别率、精度和召回率方面都明显优于 Tri-Training 算法,整体识别率提高了 7.6%,说明 Dif-TriTraining 算法针对噪声数据有较好的分类效果,具有较好的去噪能力。

实验 3　实验训练样本集中的标记比例在10% ~ 80% 的范围内,使用加噪后的辅助训练样本,抽取比例 *select_rate* 保持 0.5 不变,比较 Dif-TriTraining 算法和传统 Tri-Training 算法下的整体识别率,验证本文提出算法的有效性,结果如图 4 所示。

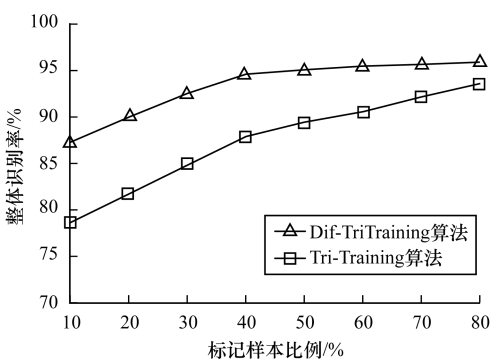

图 4　标记样本比例对整体识别率的影响

从图 4 可以看出,随着标记样本在整个训练集中所占比例的增大,Dif-TriTraining 算法和 Tri-Training 算法下的整体识别率都有所提升,但 Tri-Training 算法的性能越来越接近于 Dif-TriTraing 算法,这是因为初始已标记样本所占的比例越高,受噪声数据的影响就越小,2 种算法之间的差距会逐渐降低,然而初始标记样本所占比例越高也就意味着需要付出的代价越高。当标记所占比例在 30% ~40% 时本文算法整体识别率达到了 93%,标记比例保持在这个范围是比较合适的。

实验 4　为验证抽取比例是否会对实验结果产生影响,实验样本中的抽取比例在 0.1 ~0.8 的范围内,有标记样本所占比例为 30%,使用加噪过后的辅助训练样本所占比例为 70%,比较 2 种算法的整体识别率,在使用 Tri-Training 算法进行分析时为保证样本数量的一致性,在辅助训练样本中按相同的比例随机抽取样本用于重新训练分类器然后分析结果,实验结果如图 5 所示。

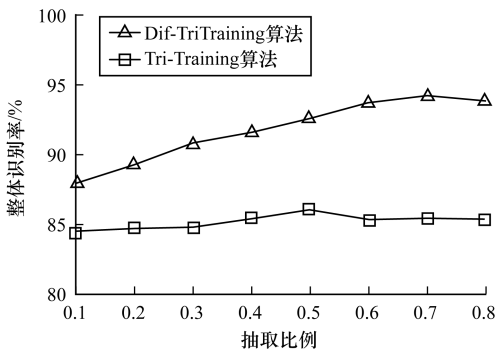

图 5　抽取比例对实验整体识别率的影响

图 5 比较 2 种算法在新加入标记样本数对整体识别率的影响,可以看出,本文 Dif-TriTraining 算法影响更加明显,而传统的 Tri-Training 算法的变化很小,使用 Dif-TriTraining 算法后在 0.1 ~0.7 范围内整体识别率有不同程度的增长,超过这个范围后开始削弱,这说明一开始引入大量的标记正确的样本重新训练分类器能够提升分类器效果,但当 *select_rate* 过大时会把大量的噪声数据引入反而会削弱噪声过滤的效果。

5　结束语

为提高空间信息网络的可管性和可控性,实现资源的合理分配和业务的高效编排,本文提出一种软件定义空间信息网络业务识别技术。根据空间信息网络的实时性要求实现一种在线业务流量识别方案,并针对传统协同训练和空间信息网络特征容易引入噪声等问题,提出一种具有噪声过滤功能的协同训练算法。实验结果表明,该算法具有良好的分类效果。

参考文献

[1] YAN J, YUN X, WU Z. Online traffic classification based on co-training method[C]//Proceedings of the 13th International Conference on Parallel and Distributed Computing, Applications and Technologies. Beijing, China:[s. n.],2012:125-136.

[2] SUN M,CHEN J. Research of the traffic characteris-tics for the real time online traffic classification[J]. The Journal of China Universities of Posts and Telecommunications,2011,18(3):92-98.

[3] 彭立志. 互联网流量识别研究综述[J]. 济南大学学报(自然科学版),2016,30(2):95-104.

[4] 王丽冲,姚秀娟,闫毅. 一种基于 OpenFlow 的软件定义卫星网络架构设计方案[J]. 电子设计工程,2016,24(17):85-89.

[5] 杨力,孔志翔,石怀峰. 软件定义空间信息网络多控制器动态部署策略[J]. 计算机工程,2018,44(10):58-63.

[6] NAZARI S,DU P,GERIA M,et al. Software defined naval network for satellite communications[C]//Proceedings of 2016 IEEE Military Communications Conference. Washington D.C.,USA:IEEE Press,2016:124-136.

[7] ELMASRY G,MCCLATCHY D,HiINRICH R,et al. A software defined networking framework for future airborne connectivity[C]//Proceedings of ICNS'17. Herndon,USA:IEEE Press,2017:215-224.

[8] AMARAL P,DINIS J,PINTO P,et al. Machine learning in software defined networks:data collection and traffic classification[C]//Proceedings of the 24th International Conference on Network Protocal. Singapore:[s. n.],2016:321-329.

[9] BERNAILLE L,TEIXEIRA R,AKODKENOU I,et al. Traffic classification on the fly[C]//Proceedings of ACM SIGCOMM'06. New York, USA: ACM Press, 2006:158-167.

[10] ESTE A, GRINGOLI F, SALGARELLI L. On the stability of the information carried by traffic flow features at the packet level[J]. ACM SIGCOMM Computer Communication Review,2009,39(3):13-18.

[11] 赵树鹏,陈贞翔,彭立志. 基于流中前 5 个包的在线流量分类特征[J]. 济南大学学报(自然科学版),2012,26(2):156-160.

[12] 刘珍,王若愚,蔡先发,等. 互联网流量分类中流量特征研究[J]. 计算机应用研究,2017,34(1):8-13.

[13] 柏骏,夏靖波,吴吉祥,等. 实时网络流量分类研究综述[J]. 计算机科学,2013,40(9):8-15.

[14] BERNAILLE L,TEIXEIRA R,SALAMATIAN K. Early application identification[C]//Proceedings of CoNEXT'06. New York,USA:ACM Press,2006:147-158.

[15] 彭建芬,周亚建,王枞,等. TCP 流量早期识别方法[J]. 应用科学学报,2011,29(1):73-77.

[16] MOORE A W, ZUEV D. Internet traffic classification using Bayesian analysis techniques[C]//Proceedings of International Conference on Measurement and Modeling of Computer Systems. Alberta, Canada:[s. n.],2005:50-60.

[17] WEI L, ABDIN K, DANN R, et al. Approaching real-time network traffic classification[D]. London, UK: London University,2006.

[18] 张剑,钱宗珏,寿国础,等. 在线聚类的网络流量识别[J]. 北京邮电大学学报,2011,34(1):103-106.

[19] 郭翔宇,王魏. 一种改进的协同训练算法:Compatible Co-training[J]. 南京大学学报(自然科学版),2016,52(4):662-671.

[20] ZHOU Z H,LI M. Tri-Training:exploiting unlabeled data using three classifiers[J]. IEEE Transactions on Knowledge and Data Engineering,2005,17(11):1259-1541.

[21] 徐鹏,林森. 基于 C4.5 决策树的流量分类方法[J]. 软件学报,2009,20(10):2692-2704.

编辑 索书志

(上接第 17 页)

[6] 狄浩. 虚拟网络的高效和可靠映射算法研究[D]. 成都:电子科技大学,2013.

[7] DI H,YU H,ANAND V,et al. Efficient online virtual network mapping using resource evaluation[J]. Journal of Network and Systems Management,2012,20(4):468-488.

[8] 刘彩霞,李凌书,汤红波,等. 基于子图同构的 vEPC 虚拟网络分层协同映射算法[J]. 电子与信息学报,2017,39(5):1170-1177.

[9] FENG M,LIAO J X,WANG J Y,et al. Topology-aware virtual network embedding based on multiple characteristics[C]//Proceedings of 2014 IEEE International Conference on Communications. Washington D.C.,USA:IEEE Press,2014:2956-2962.

[10] 李振涛,孟相如,赵志远,等. 基于节点抗毁能力感知的虚拟网络可靠映射算法[J]. 计算机工程,2017,43(9):62-67.

[11] 龚水清,陈靖,黄聪会,等. 信任感知的安全虚拟网络映射算法[J]. 通信学报,2015,36(11):180-189.

[12] 赵志远,孟相如,苏玉泽,等. 基于节点邻近感知与路径综合评估的虚拟网络映射算法[J]. 电子与信息学报,2017,39(8):1979-1985.

[13] KELLY D J,COMM B,O'NEILL G M. The minimum cost flow problem and the network simplex solution method[D]. Dublin, Ireland: University College Dublin,1991.

[14] MURAMATSU M. On network simplex method using the primal-dual symmetric pivoting rule[J]. Journal of the Operations Research Society of Japan,2017,43(1):149-161.

[15] WANG I L,LIN S J. A network simplex algorithm for solving the minimum distribution cost problem[J]. Journal of Industrial and Management Optimization,2017,5(4):929-950.

[16] ZHANG L X, DEBORAH E, JEFFREY B, et al. Named Data Networking(NDN) project[J]. Journal of the Transportation Research Board,2010,1892(1):227-234.

编辑 刘盛龄