一种基于 SVM 的 VoIP 网关节点识别方法*

梁 伟, 陈福才, 李邵梅
(国家数字交换系统工程技术研究中心, 郑州 450002)

摘 要: 针对互联网中常见的 VoIP 语音网关,重点从呼叫及语音数据流特征的选取出发,提出了一种基于 SVM 的 VoIP 网关节点识别方法,并利用真实网络数据对其进行了测试。结果表明,该方法能使待识别数据的时段规律和周期性得到较好体现,识别准确率高,适用于对不同 VoIP 的语音网关节点进行识别。

关键词: VoIP; 支持向量机; 网关节点
中图分类号: TP393 **文献标志码**: A **文章编号**: 1001-3695(2012)12-4706-05
doi:10.3969/j.issn.1001-3695.2012.12.079

Method of VoIP MediaGateway identification based on SVM

LIANG Wei, CHEN Fu-cai, LI Shao-mei
(*National Digital Switching System Engineering & Technological Research Center*, *Zhengzhou* 450002, *China*)

Abstract: The paper proposed a novel method to identify several familiar VoIP MediaGateway based on SVM, which was mainly by select the VoIP calls and VoIP flow features of classification. It tested the true Internet data. The experimental results show the method is adapted well for data period feature of VoIP, achieves high indentification accuracy for the Internet data including many different type of VoIP MediaGateway.

Key words: VoIP(voice over Internet protocol); SVM(support vector machine); MediaGateway

VoIP 业务成本低廉、部署方便,在全球范围内深受非基础电信运营商的青睐,在话音通信业务中的比例也持续上升。互联网上存在的 VoIP 服务通常可以分为以下几种形式,即 PC to PC、PC to phone、phone to PC。其中,PC to phone、phone to PC 的 VoIP 服务需要 VoIP 网关作为其互联网与 PSTN 网的连接点。通话过程中,信令和媒体由一种网络形态进入另一种网络形态时,VoIP 网关为其提供转换功能,以达到互通目的。图 1 为典型的 VoIP 网关连接示意图。

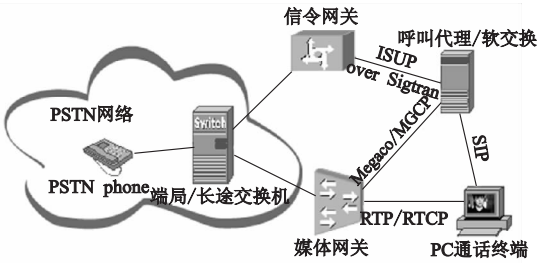

图1 VoIP网关连接示意

作为互联网技术面向传统电信业务的一种应用,VoIP 继承了互联网的开放性、灵活性,但也将互联网的安全隐患带入到电信业务领域。特别是近年来,源自互联网,经由非法、隐性运营的 VoIP 网关进入电信网的广告和欺诈类电话越来越多,给人民群众的日常生活造成了诸多干扰和危害,因此对 VoIP 进行合理管控显得尤为重要。作为 VoIP 业务管控的关键技术之一,VoIP 网关节点识别技术意在通过对存在于互联网中的 VoIP 网关节点进行辨识,并为后续对通过该节点的 VoIP 话务进行管控提供重要支撑,因此成为继 VoIP 流量识别技术之后又一迫切需要研究解决的课题。

从 IP 地址的范围看,互联网遍布的网络节点难以数计,节点类型更是细小繁多,根据研究对象的不同,将涉及到 VoIP 语音数据包通信的节点分为类 PC 用户(通常不表现出语音数据包的网络汇聚现象)、VoIP 网关(开放协议 VoIP 网关、私有协议 VoIP 网关)这两大类节点。鉴于管控的必要性,需要从众多 VoIP 媒体数据包涉及到的 IP 地址中识别出属于 VoIP 网关的地址。本文拟针对通过该节点的一段时间内的 VoIP 媒体数据包的某些规律和特点,借鉴 SVM 的方法,判断该节点是否属于 VoIP 网关。如无特殊说明,本文研究所指的均为 VoIP 媒体网关(MG)。

1 互联网节点识别现状

目前,公开文献中互联网节点识别方法多集中在关于 P2P 节点的识别方法,主要有以下四类。

1.1 基于明文特征的节点识别方法

Saroiu 等人[1]从主动测量 P2P 节点出发,首先使用了主动爬虫技术,对 Napster 和 Gnutella 这两种 P2P 应用进行了测量研究。Sen 等人在 2004 年提出一种基于特征码的 P2P 流量识别方法,根据 P2P 应用的特征码及其在数据包中的偏移位置以识别 P2P 流量[2]。该方法对于已知特征字段的 P2P 应用有一定的识别效果,但是对于未知的应用或者应用字段不明显的情况则无能为力。

1.2 基于传输行为特征的节点识别方法

Wang 等人[3]提出,由于 P2P 节点的双重特性,其既具有发起连接又具有回应连接,认为与某一 P2P 节点通信的其他

收稿日期: 2012-03-29; **修回日期**: 2012-05-23 **基金项目**: 国家"863"计划资助项目(2011AA010605)

作者简介: 梁伟(1984-),男,硕士研究生,主要研究方向为电信网信息安全(lianglevis@163.com);陈福才(1974-),男,研究员,主要研究方向为电信网信息安全;李邵梅(1982-),女,讲师,主要研究方向为电信网信息安全.

P2P 节点中不同的 IP 数与端口数大致相等,据此提出针对基于 TCP 的 P2P 节点识别方法。Lu 等人[4]则认为 P2P 节点的双重特性导致 P2P 节点下载的数据将会被用于上传,即 P2P 节点中下载的数据与上传的数据具有公共部分,因此可被用来识别节点性质。文献[5]中连接成功率(connection success rate)的概念为文献[6]对 BitTorrent 及 Gnutella 主机的识别提供了重要参考,其通过三项 P2P 固有的行为特征来进行 P2P 节点识别。该方法抓住 P2P 应用的行为规律进行判断,应用简单,有着较高的识别速度,但由于并不深入协议内部特征,导致误判率较高。

1.3　基于统计模型的节点识别方法

Erman 等人[7]采用建模的方法研究了 BitTorrent 的连接长度、连接持续时间、连接到达时间间隔、分片的下载时间等特征,利用 Gaussian、Weibull、Hyper-exponential、Log-normal 等分布以及其组合分布进行拟合分析。文献[8]通过测量 Edonkey、Gnutella、FastTrack、MP2P 及 OpenNap 五种 P2P 应用,分析了用户行为特征及每个用户文件数的变化规律,并建立了相关的模型。文献[9]针对类 BitTorrent 应用提出一种基于远端主机离散度(remote hosts discreteness,RHD)的 P2P 节点识别方法。该方法需要大量数据来建立统计模型,识别时需要的节点数据量相应较多,虽然操作简单,能够反映节点的长时规律,但识别速度较慢,对于当前数据的识别有一定的延迟。

1.4　其他节点识别方法

文献[10]提出一种根据节点关联的网络拓扑直径进行 P2P 节点识别的方法。根据“P2P 系统所形成的网络比其他应用具有更大的直径”这一事实,通过记录网络节点间的连接状况进而得到 P2P 系统的连接拓扑图。若网络直径超过某个门限值,那么该节点就标记为 P2P 节点。文献[11]基于 P2P 节点同时采用 TCP 及 UDP 进行通信、P2P 节点的随机端口、P2P 节点较少访问 DNS 服务器等特征,采用基于柔性神经树(flexible neural tree,FNT)算法进行 P2P 的识别。上述方法都是根据 P2P 应用中一些较强的特征进行判别,但这些特征参数并不充分,因此虽然可以达到一定的识别效果,但漏判率及误判率指标难以满足应用需求。

从公开文献来看,当前与 VoIP 网关节点识别技术直接相关的资料并不多见,以上分析主要是试图从当前大量的关于 P2P 的研究中找出有启发意义的方法。其中,关于用户行为特征和统计建模的方法对 VoIP 网关节点识别有一定启示作用。本文针对 VoIP 呼叫的流特征在单位时间上较强的阶段性和周期性特点,采取多维特征描述其规律,利用 SVM 分类算法来实现对 VoIP 网关节点的识别。

2　VoIP 网关节点特征分析

2.1　VoIP 呼叫特点分析

1)VoIP 呼叫的媒体数据包包长特点

互联网中的 VoIP 语音数据包是由一些公有及私有 VoIP 信令协议引导的媒体数据包,不同的 VoIP 信令协议采用不同的语音编码标准。如采用 SIP、H.323 等公有协议的 VoIP 应用所采用的语音编码标准通常为 G723、G729、G711 等,而诸如 QQ、Skype 等私有协议采用的语音编码标准则各不相同。不同的语音编码标准决定了语音数据包的长度。研究国际链路中 VoIP 呼叫的包长分布[12],发现使用 G723、G729 的普通 VoIP 的语音数据包长通常以 77、82、106 Byte 为主,而 Skype 语音包长比较分散,基本上连续分布在 74 ~ 300 bit 之间,其中以 74 和 154 bit 的语音包为主。

2)VoIP 呼叫的周期性和阶段性特点

从 VoIP 呼叫分布的周期性特点来看,与国际 PSTN 话务类似,数据在观测时间为 24 h 和 168 h 的时间点上表现出很强的周期性。由于 VoIP 网关节点单位时间内的呼叫总量通常维持在一定的水平之上,根据租用其相对固定的用户群体和通话习惯得知,以一天或一周为单位,其呼叫流量的各项统计特征表现出较强的周期性。

从通话时长的分布[12]来看 VoIP 呼叫阶段性特点,工作时段平均通话时长为 199 s,而非工作时段平均通话时长为 350 s。VoIP 忙时的平均时长为 81 s,闲时 169 s,且小于 2 min 的通话时长占 80%。工作时段的 VoIP 呼叫量要明显大于 8:00 前和 17:30 之后的时间段。在传统忙时里,通话时长短,通话频率较高;而传统闲时里,通话时长较长,频率较低,话务模型更近似于 PSTN。

若要详细刻画网络中某个节点的 VoIP 呼叫特点,须采用能够反映周期性与阶段性两个特点的识别特征。

3)VoIP 业务模式的特点

不同的 VoIP 业务有着不同的网络模式。如 QQ 通话,由于其现阶段还不具有与 PSTN 电话和移动电话通话的能力,所以其通话的模式局限于 PC 到 PC 的模式。如 Skype 通话,作为全球流行范围最广的 VoIP 服务提供模式,其具备 Skype-Skype、Skype-phone、phone-Skype 诸多模式。如 SIP、H.323 电话,其呼叫的模式包括 PC-phone 或 phone-PC,采用专有 VoIP 网关进行通信,也包括终端均为 IP phone 的 phone-phone 的形式,呼叫的模式其实相当于 PC-PC,也包括 phone-phone 的电话卡类的业务形式。因此,不同的网络节点之中,使用 VoIP 业务存在差异性,这种差异性可以用来作为判断节点类型的因素。

2.2　VoIP 网关的特点分析

为了方便下面的叙述,首先对两类节点进行说明。在互联网中参与语音通信的网络节点主要分两类:a)像 PC 一样主要是使用 VoIP 终端软件进行语音通信的节点,这类节点既包含常见的 PC,又包含一些其他类型的主机,如使用语音类终端进行通信的服务器,这些节点在本文的研究中统一被视为类 PC 节点;b)参与 VoIP 语音通信的媒体网关(media gateway, MG),也就是本文所说的 VoIP 网关,这是本节重点要研究的识别对象。本文研究的重点就是要把类 PC 节点和 VoIP 网关区分开。

1)类 PC 节点特点分析

对于 PC 节点而言,其进行 VoIP 通信,绝大多数情况下,同一时刻发起的会话只有一个,极少情况下会出现多个会话同时进行,如多人会话状态或者多个终端同时进行会话。因此会话过程中,通过该节点的媒体流量在网络未发生拥塞等异常情况下,媒体数据包严格服从相应的语音编码标准,呈现出一定的特征。

2)VoIP 网关节点特点分析

作为传送 VoIP 呼叫的重要设备,VoIP 网关具有功能单一性的特点,其流量具有以下特征:

a)通过该节点的媒体数据流无一例外均为语音数据包。

b)因为一定的网关只支持某种或某几种语音编码标准,

所以经过成功转发的语音数据包无一例外均为已知的几种编码中的某一种。

c)由于网关通常会同时处理来自多路的会话,因此其表现最明显的就是并发流的特性,与其通联的 IP 地址上体现出较强的一对多和多对一的特性,这一点与 PC 到 PC 的通话模式有较大区别。具体对比如图 2 所示。

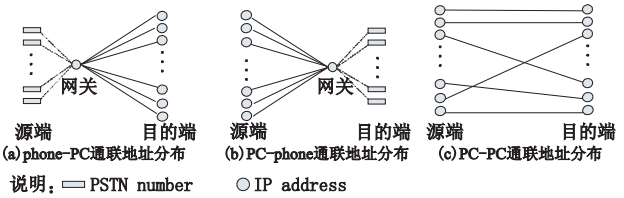

图2 几种VoIP呼叫方式的地址通联模式差异

综上所述,可以通过分析网络中节点的流量特征进行 VoIP 网关的识别。

2.3 分类特征定义及选取

结合以上分析,本文采用以下四个特征对 VoIP 网关节点进行识别:

a)流速(Flrate)。本文以每小时流经节点的数据包总数来衡量。从 VoIP 网关一对多的通联特性易知,单位时间内通过网关的数据包的总数应该在数量级上区别于普通的 PC 用户。

b)数据包长度变化范围(Lenrange)。采用不同的语音编码标准意味着不同的数据包长。对于 VoIP 而言,有的数据包采用固定包长语音编码标准,有的采用可变包长语音编码标准。从一些观察数据可知,多数网关支持某几种固定语音编码标准,而采用可变长语音编码标准的某些 VoIP 应用更多地出现在 PC 用户中。设置这样一个特征可以很好地表征单位时间内是否出现了不同于固定包长的语音数据包。本文所提及的数据包长度均指从 IP 头部开始计算的数据长度,以字节为单位。通常来讲,网关的该项指标较为固定,而类 PC 节点则表现出不稳定性。

c)包长类型数(Lentypeno)。本指标是指待分析时间段内,流经节点的数据包拥有的包长类型的个数。该指标主要与流经节点的会话数以及会话数采用的编码标准有关。

d)并发连接比例(Interatio)。在介绍这个指标之前,先来看几个概念。

(a)连接。类似利用五元组定义流,将相同五元组的、以该节点地址为目的 IP 的单向语音数据包的、符合一定条件的集合定义为连接。

(b)生存期。某个连接从开始到结束的这段时间称之为该连接的生存期。

(c)单一连接和并发连接。对上述的每一个连接定义两个状态。第一个状态是该连接在自己生存期内始终单独出现在某节点,期间没有出现别的关于该节点的连接,这种状态记为单一连接。相反,第二个状态是指在这条连接的生存期内,另外出现了至少一个关于该节点的连接,则这个状态记做并发连接。

并发连接比例指的是一段时间内,并发连接的数量占总连接数的比值,即并发连接/(并发连接 + 单一连接)。当该节点同时与多个 IP 地址进行 VoIP 通信时,并发连接的数量会明显增多。在固定时长内, VoIP 网关的会话并发连接比例和类 PC 节点的会话并发连接比例。这里的固定时长通常以 ms 为单位。

综合本节定义的上述特征,本文采用下列特征组对网络中节点 k 在时段 i 内的特性进行描述:$\text{flowDescriptor}^k[i] = \{\text{Flrate}[i], \text{Lenrange}[i], \text{Lentypeno}[i], \text{Interatio}[i]\}$。

进一步结合上节中分析的 VoIP 流量分布在时间上的规律性,不同时段内的 flowDescriptor$[i]$ 在数值上应该有较大的差异。考虑这种时间特性,采用天为单位对每个节点的流量进行描述,得到网络中每个节点 k 的 VoIP 流量特征:$\text{FullFlowDescriptor}[k] = \{\text{flowDescriptor}^k[0], \text{flowDescriptor}^k[1], \cdots, \text{flowDescriptor}^k[23]\}$。这样,每个节点的流量特征维数为 4×24。

该特征的维数较高,无法用传统的基于线和平面的方法对其分类,所以本文采用基于机器学习的分类方法。因为 VoIP 网关的识别本质上是一个二分类问题,并且实际应用中可以事先获取的 VoIP 网关的数据有限,在这种情况下,本文采用更适于在小训练样本的情况下进行分类的 SVM 算法。

3 基于 SVM 的 VoIP 网关节点识别方法

3.1 SVM 原理简析

支持向量机[13,14]是 Vapnik 于 1995 年首次提出的,它是建立在统计学习理论基础上的机器学习方法。SVM 本质上是寻找最大分类间隔的线性分类器,根据样本的情况分为线性可分、线性不可分和非线性问题三类。线性可分情况下对应着最大分类间隔算法,线性不可分情况下对应着软分类间隔算法,非线性问题通过核函数转换为线性问题解决。

SVM 的分类判决函数形式为

$$f(x) = \sum_{i=1}^{N} \alpha_i y_i (x \cdot x_i) + b \tag{1}$$

其中:x_i 为支持向量;$y_i \in \{+1, -1\}$ 是相应的类别标签;N 是支持向量的个数;α_i 和 b 是通过训练数据优化得到的数值。

SVM 是从线性可分情况下提出的最优分界面。用于非线性分类问题时,通常先把分类的数据投影到高维的线性可分空间,然后在高维空间中可以通过非线性变换转换为某个高维空间中的线性问题,在变换空间求线性的最优分类面。

在实际应用中,这种投影通过核函数来实现,核函数在支持向量机的理论体系中占有非常重要的地位。根据泛函的有关理论,核函数 $K(x_i, x_j)$ 满足 Mercer 条件,它对应某变换空间中的内积:

$$K(x_i, x_j) = \varphi(x_i) \cdot \varphi(x_j) \tag{2}$$

最终,SVM 的判决函数转换为如下形式:

$$f(x) = \sum_{i=1}^{N} \alpha_i y_i K(x, x_i) + b \tag{3}$$

3.2 基于 SVM 的 VoIP 网关节点识别模型设计

如上所述,本文研究的重点就是基于各网络节点的流量特性,区分类 PC 节点和 VoIP 网关。由此拟基于 SVM 构建 VoIP 网关识别模型,如图 3 所示。

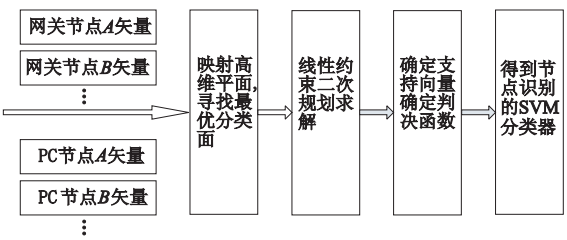

图3 VoIP网关识别模型的建模流程

在建模阶段,首先收集多个 VoIP 网关的流量数据,按照 2.3 节的方法提取各个网关的 VoIP 流量特征 FullFlowDescriptor[k](k=0,1,…,m),m 是 VoIP 网关的个数;然后收集多个 PC 节点的流量数据,按同样的步骤提取各个节点的 VoIP 流量特征 FullFlowDescriptor[k](k=0,1,…,n),n 是 PC 节点的个数;再把这两组特征和它们对应的类别标签(第一组的标签为 1,表示是 VoIP 网关;第二组的标签是 0,表示不是 VoIP 网关)送入 SVM 的训练模型,得到模型的参数。

在识别阶段,对于要识别的网络节点,同样采集其流量数据,提取 VoIP 流量特征 FullFlowDescriptor,然后送入 VoIP 网关识别模型进行判决,如果模型输出是 1,就代表该节点是 VoIP 网关;否则就是 PC 节点。

上述 SVM 模型的建模和判决都借助开源平台 Weka[15] 完成。

3.3　基于多时段的 SVM 识别模型设计

本文对每个节点的 VoIP 流量特征的描述是基于天的,从 2.1 节的分析可知,VoIP 呼叫的数目在工作日和节假日有较大的差异。针对这一特点,分别建立工作日的 VoIP 网关模型 A 和节假日的 VoIP 网关模型 B。如果采集到的待识别的节点的流量数据是来自工作日的,将其送入 A 分类器进行分类;否则,将其送入 B 分类器进行分类,如图 4 所示。

关于两个分类器结果的抉择,本文采取的策略是,当同一网络节点的工作日数据或者节假日数据有一方被识别为网关节点,鉴于网关节点相对于类 PC 节点的微量性和分类器的准确性,则该节点被标记为网关节点。

4　实验分析

4.1　评价指标

本文把网络中的通信节点分为 VoIP 网关(MG)和类 PC 两大类。对这两类节点的分类结果可以得到如表 1 的混淆矩阵(confusion matrix)。

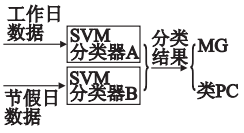

图4　分时段SVM分类技术路线

表1　分类器的混淆矩阵

分类前	分类后	
	MG	类PC
MG	TP	FN
类PC	FP	TN

TP(true positive):真实类型为网关,被正确识别为网关的样本数量。

TN(true negative):真实类型为类 PC,被正确识别为类 PC 的样本数量。

FP(false positive):真实类型为类 PC,被错误识别为网关的样本数量。

FN(false negative):真实类型为网关,被错误识别为类 PC 的样本数量。

与识别性能相关的主要有以下几个评价指标:

$$\text{召回率(recall)}\quad \text{recall}=\text{TP}/(\text{TP}+\text{FN}) \tag{4}$$

$$\text{精度(precision)}\quad \text{Precision}=\text{TP}/(\text{TP}+\text{FP}) \tag{5}$$

$$\text{整体准确率(overall accuracy,OA)}\ \text{overallAccuracy}=(\text{TP}+\text{TN})/(\text{TP}+\text{FN}+\text{FP}+\text{TN}) \tag{6}$$

4.2　实验数据

本文的实验数据采自某系统,现已知该网络范围内有 VoIP 网关(开放协议 VoIP 网关、私有协议 VoIP 网关)节点共计 17 个,其中开放协议 VoIP 网关 10 个,私有协议 VoIP 网关 7 个。已知存在的 VoIP 应用包括开放协议 SIP、H.323 引导的网络电话,及私有网络电话 Skype、Googletalk、QQVoice、UUcall、KC 网络电话、阿里通网络电话。现将由该网络中采集、整理的语音数据 NET_SET 共 20 GB(时间跨度为一周)用于 SVM 分类器的训练。该实验数据中,通过 17 个网关节点的语音数据约为 9 GB,即属于 PC-phone 通话模式的数据约为 9 GB,属于 PC-PC 通话模式的数据约为 11 GB。所有的数据均涉及到拥有公网 IP 的类 PC 节点,或发起呼叫,或接听呼叫。

通过对 9 GB 数据的预处理,得到以 17 个网关节点为目的 IP 的单向 VoIP 语音数据包约 4.5 GB,此作为分类正向训练数据。通过对 20 GB 数据以类 PC 节点为目的 IP(除 17 个专属网关节点之外的其他 IP)进行预处理,得到单向语音数据包约 15.5 GB,此作为分类反向训练数据。反向训练数据通过从与 17 个网关节点通信的类 PC 节点中有应用代表性的进行选择。

为符合分类算法要求,将数据集 NET_SET 按照时间分为两个数据子集,将 NET_SET 分为 NET_SET1 和 NET_SET2,NET_SET1 用来训练分类器,NET_SET2 用来测试分类效果。

4.3　数据预处理

首先对已经分拣出的涉及 VoIP 通信的语音数据包进行单向整理,即将相同目的 IP 地址的数据包检出并根据时间先后顺序进行存储。为了体现 VoIP 呼叫与传统 PSTN 呼叫相似的规律性和周期性特点,以该样本节点一天的语音数据包量作为特征提取的基本数据。特征由如下方法获取,按照 0-1,1-2,2-3……这样的单位小时时间段划分一天的数据包,分别对每段数据进行统计得出需要的四个特征值,构成 flowDescriptor^k[i],则该样本节点全天的数据可以得到 4×24 维的特征值,以此来描述该样本本身。

关于计算一段时间内并发连接的数量,先统计该时间段内该节点总共出现的连接总数,当然是指之前定义的以某个 IP 地址为目的地址的连接,并同时记录该条连接生存期。将一条连接标记为并发连接后,将与其时间交叉的所有连接都标记为并发连接,然后将其死亡时刻之后的开始时间最早的一条连接进行判断,如遇别的生存期与之有交叉的将其判断为并发连接。如此重复以上步骤来判断并对符合条件的连接数进行计数。如图 5 所示,当前 L_1,L_2……均为一段时间内流向目标节点的单向数据连接,假设当前以 L_1 是否为并发连接的判别对象,当标记 L_1 为并发连接时,易见 L_4、L_5、L_6 均与 L_1 有时间交叠,也将其标记为并发连接,之后,由于 L_2 比 L_3 更接近于 L_1 结束的时间,因此从 L_2 开始判断,如此反复。图 5 中,L_7 是单一连接。

4.4　实验平台

实验采用新西兰怀卡托大学 Witten 教授等人基于 Java 编程开发的开源工作平台 Weka-3.6[15],该平台实现了朴素贝叶斯、决策树、贝叶斯神经网络、SVM 等多种机器学习方法。本文进行分类的实验环境为一台普通 PC,系统配置为 Windows XP 操作系统,CPU 为 Intel Pentium4 2.8 GHz,内存为 DDRⅡ 2 GB,分类实验进行时没有其他应用程序运行。

4.5　实验项目及结果分析

1)识别性能分析

a)工作日数据实验的结果分析

为了验证方法在不同样本集规模下对识别的影响,样本个

数在10~35个之间变化,每次变化增加5个样本,同时采用了其他三种方法进行了相同样本集规模下的对比。结果如图6所示。

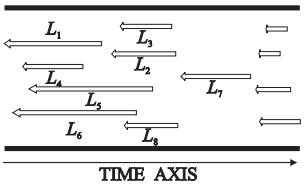

图5 并发连接与单一连接

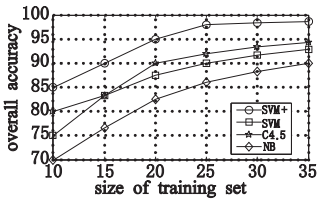

图6 几种方法在工作日不同样本集下的分类效果

图6中标签为SVM+的方法是本文所提的方法,鉴于样本数量有限,而且特征维数较高,以便于正常使用以下几种机器学习算法,图中标签为SVM, C4.5, NB的方法采用的特征是对一天的数据量不分时段、集中提取总的流速、包长范围、包长类型数、并发连接比例。横坐标代表训练样本集规模,纵坐标代表识别的整体准确率。

随着训练样本的增加,四种方法的识别准确率也都随之提升。在同样的训练样本集情况下,识别整体准确率由高到低依次为SVM+ > C4.5 > SVM > NB。容易看出,SVM+在每个训练样本集下,识别准确率都明显优于其他三种。该方法将一天的数据按单位小时长的时段划分,较好地反映了一天中通话忙闲时的规律性与周期性,继而通过提取各段数据的特征来对分类器进行训练。实验的准确率达到了预期效果。

当训练样本集规模为35时,本文对四种方法总体识别准确率、精度和召回率进行比较,如表2所示。

表2 NET_SET在准确性指标上的实验结果

算法	OA/%	recall/%	precision/%
SVM+	98.57	100.00	97.22
SVM	92.85	94.29	91.67
C4.5	94.29	94.29	94.29
NB	90.00	91.43	88.89

易见SVM+在当前情况下召回率达到了100%,即说明对于专属VoIP网关的识别没有出现漏判,精度反映出只是出现了一个误将类PC用户识别为网关节点的情况。总体而言,SVM+各项具体指标均优于所列其他方法,达到了使用需求。

b)节假日数据实验的结果分析

节假日数据实验方法同工作日数据,得到的结果如图7所示。图7中,节假日数据样本在进行分类时,相比较工作日数据的表现来说,出现了较多的重合点,平滑度上也易见不及工作日数据的表现,除了本文所提的方法效果较其他方法比较明显之外,其他三种方法表现变化较大。这说明,节假日的数据有区别于工作日的独有特点。这与用户的通话行为习惯有着较大的关系。经前面VoIP呼叫特点分析可知,在非工作日里,私人通话时间增多,但呼叫频率降低了,单个呼叫时长增加了;工作性质的电话减少,呼叫频率同时降低,单个呼叫时长没有明显变化。总体来看,这说明非工作日的话务量减少,单个呼叫时间变长,呼叫频率下降。由此导致数据总体发生变化,通过网关的数据量减少,而通过类PC用户的数据量也减少了,这样就导致四个参数中流速变小,数据包长度变化范围变小,包长类型数变小,并发连接比例变小,从而整体上影响了模型构建及分类的效果。

根据表3所示与前面在工作日里的分类结果相比,在训练样本数均为35的情况下,各项指标与之相差无几,说明了样本数量在超过节点个数两倍的情况下,两种时间段内数据的分类表现是一致的,也说明了样本规模超过30时,训练所得的分类器效果是明显的,表现出一定的稳定性。

表3 节假日数据在准确性指标上的实验结果

算法	OA/%	recall/%	precision/%
SVM+	98.57	100.00	97.22
SVM	92.85	94.29	91.67
C4.5	94.29	94.29	94.29
NB	91.43	91.43	91.43

2)特征比较分析

为了探讨所定义及选取的特征对分类具体的贡献,本文利用排除法的原理通过在减少某项特征的情况下进行训练和识别的方法来分析各种特征对识别的贡献。实验结果如图8所示。

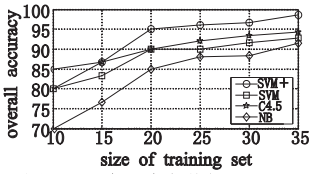

图7 几种方法在节假日不同样本集下的分类效果

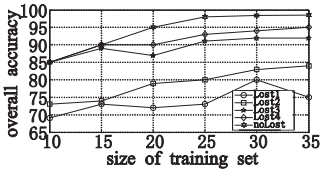

图8 单一特征缺失后分类效果

图8中标签Lost1、Lost2、Lost3、Lost4分别代表唯一缺少并发连接比例、缺少流速、缺少包长范围、缺少包长类型数这四种情况,noLost代表不缺少任何一种特征的情况。从缺失后的情况看来,并发连接比例的缺失直接导致分类整体准确率下降至80%以下,维持在65%~75%,并且出现了较大的波动,分类效果不稳定。紧随其后影响较大的一个特征是流速,这是一个最直观且易计算的特征,但直接反映了网关节点与类PC节点在语音通话数量上的区别,因此也是一个重要的特征,优先级要高于包长范围和包长类型数。这样得到四个特征对于分类贡献率的大小关系:并发连接>流速>包长范围>包长类型数。

5 结束语

VoIP网关识别是互联网业务感知及监管中继VoIP流量识别技术之后的又一关键技术。目前针对网关节点识别的研究过多集中于对P2P节点的识别,而针对VoIP网关识别的研究并不多见。本文提出了一种针对VoIP网关进行识别的方法,通过选取有效反映流经VoIP网关的数据规律的几个特征,结合SVM算法,取得了较好的识别效果。由于没有充分考虑提取特征时计算复合特征所带来的时间消耗,该识别方法用于离线对VoIP网关识别是可行的,但是否适用于在线识别还有待进一步研究。

参考文献:

[1] SAROIU S, GUMMADI P K, GRIBBLE S D. A measurement study of peer-to-peer file sharing systems[D]. San Jose: The International Society for Optical Engineering, 2002:156-170.

[2] SUH K, FIGUEIREDO D R, KUROSE J, *et al*. Characterizing and detecting skype-relayed traffic[C]//Proc of the 25th IEEE International Conference on Computer Communications. Barcelona: IEEE, 2006: 1-12.

[3] WANG Jin-song, ZHANG Yan, WU Qing, *et al*. Connection patten-based P2P application identification characteristic[C]//Proc of Network and Parallel Computing Workshops. Washington DC: IEEE Computer Society, 2007:437-441.

(下转第4713页)

之比,达到了预期的效果。

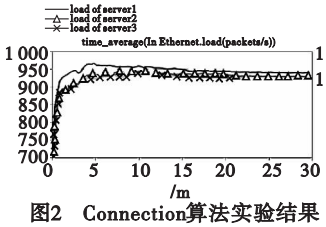

图2　Connection算法实验结果

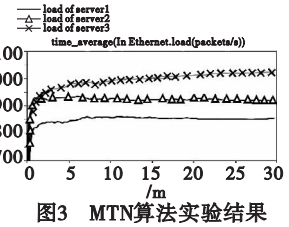

图3　MTN算法实验结果

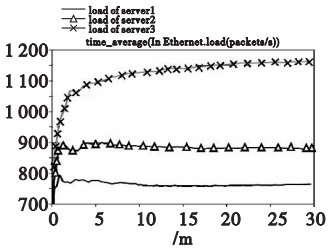

图4　Weight算法实验结果

负载均衡算法最终的目的是充分利用节点负载和性能的信息,将任务分配到集群节点上,使得各个节点的负载与其性能成正比[9,10],为此引入了均衡效率对比三种算法的性能。均衡效率 $H(S_i)$ 是指节点实际的平均负载与按照算法分配理论上的平均负载的比值。为了简化计算步骤,取集群中的第一个节点作为参考系,使用式(10)计算。

$$H(S_i)=\frac{L(S_i)/L(S_l)}{C(S_i)/C(S_l)}\quad i=0,1,\cdots,n-1 \tag{10}$$

由均衡效率定义可知,比值越接近于 100% 算法性能越好。三种算法的均衡效率结果如表 1 所示。

表1　三种负载均衡算法的均衡效率对比

算法	服务器 S_1	服务器 S_2	服务器 S_3
Weight	100%	77.04%	77.39%
MTN	100%	72.42%	61.63%
Connection	100%	65.51%	51.44%

据表 1 可得,最小连接法的均衡效率最低,性能最高的节点均衡效率仅 51.44%,均衡效果最不理想;基于动态反馈的负载均衡算法根据不同节点的性能计算分配任务的概率,均衡效果有了较大的改进,但性能最高的节点均衡效率仅 61.63%;本文提出的算法,通过综合考虑节点的性能和负载等因素指导任务的分配,使得负载分配更加均匀、合理,各个节点分配的负载比例最接近于节点性能的比例,均衡效率达到 77.39%,相对于基于动态反馈的负载均衡算法提高了 15.76%,均衡效果有了明显的改善。

4　结束语

负载均衡算法是负载均衡集群的核心部分,优秀的均衡算法能发挥集群的优势,充分利用集群的各种资源,为用户提供高质量的服务。通过分析常用负载均衡算法的优缺点,本文提出了一种基于负载权值的动态反馈负载均衡算法,该算法的主要特点是:充分应用节点的实际负载和性能两个因素,指导任务的分配,通过负载修正保证系统的稳定性。实验表明,该算法能更好地提高均衡效率,增强负载均衡集群的稳定性。

当然,本文提出的算法也存在不足,例如使用修正负载增量来抵消节点完成任务对负载造成的影响,有一定的滞后性。另外,均衡效率虽然有了一定的提高,但与理论上存在差距。在以后的实验中,将不断完善节点负载的分配和修正机制,进一步提高算法的均衡效率。

参考文献:

[1] 郭成城,晏蒲柳. 一种异构 Web 服务器集群动态负载均衡算法[J]. 计算机学报,2005,28(2):179-184.

[2] 陈伟,张玉芳,熊忠阳. 动态反馈的异构集群负载均衡算法的实现[J]. 重庆大学学报, 2010,33(2):73-78.

[3] 田绍亮,左明,吴绍伟. 一种改进的基于动态反馈的负载均衡算法[J]. 计算机工程与设计,2007,28(3):572-573.

[4] 胡志刚,张艳平. 基于目标约束的分层动态负载均衡算法[J]. 计算机应用研究,2011,28(3):1105-1107.

[5] BRYHNI H. A comparison of load balancing techniques for scalable Web servers[J]. IEEE Network,2000,14(4):58-64.

[6] 买京京,龚红艳,宋纯贺. 集群系统中的动态反馈负载均衡策略[J]. 计算机工程,2008,34(16):114-115.

[7] 刘健,徐磊,张维明. 基于动态反馈的负载均衡算法[J]. 计算机工程与科学,2003,25(5):65-68.

[8] 操惊雷,周建国,秦磊华. 基于 OPNET 的网络压力仿真[J]. 计算机工程,2009,35(23):115-117.

[9] SOKLIC M E. Simulation of load balancing algorithms: a comparative study[C]//Proc of ACM SIGCSE Bulletin. New York: ACM Press, 2002:138-141.

[10] ALTMAN E, AYESTA U, PRABHU B J. Load balancing in processor sharing systems[C]//Proc of the 3rd Internatianal Conference on Performance Evaluation Methodologies and Tools. Belgium: ICST, 2008:35-48.

(上接第 4710 页)

[4] LU Xing, DUAN Hai-xin, LI Xing. Identification of P2P traffic based on the content redistribution characteristic[C]//Proc of International Symposium on Communications and Information Technologies. [S. l.]:IEEE, 2007:596-601.

[5] MATSUDA T, NAKAMURA F, WAKAHARA Y, *et al*. Traffic features fit for P2P discrimination[C]//Proc of the 6th Asia-Pacific Symposium on Information and Telecommunication Technologies. 2005:230-235.

[6] BARTLETT G, HEIDEMANN J, PAPADOPOULOS C. Inherent behaviors for on-line detection of peer-to-peer file sharing[C]//Proc of IEEE Global Internet Symposium. Alaska:IEEE, 2007: 55-60.

[7] ERMAN D, ILIE D, POPESCU A. BitTorrent traffic characteristics [C]//Proc of International Multi-Conference on Computing in the Global Information Technology. Washington DC:IEEE Computer Society,2006: 42-49.

[8] LLORET J, DIAZ J R, JIMENEZ J M, *et al*. Public domain P2P file-sharing networks measurements and modeling[C]//Proc of International Conference on Internet Surveillance and Protection. Washington DC: IEEE Computer Society,2006: 10-16.

[9] CHENG W Q, GONG J, DING W. Identifying BT-like P2P traffic by the discreteness of remote hosts[C]//Proc of the 32nd IEEE Local Computer Networks. Washington DC: IEEE Computer Society,2007: 237-238.

[10] CONSTANTINOU F, MAVROMMATIS P. Identifying known and unknown peer-to-peer traffic[C]//Proc of the 5th IEEE International Syrnpcsium on Network Computing and Applications. Washington DC: IEEE Computer Society,2006:93-102.

[11] CHEN Zhen-xiang, YANG Bo, CHEN Yue-hui, *et al*. Online hybrid traffic classifier for peer-to-peer systems based on network processors [J]. Applied Soft Computing,2009,9(2):685-694.

[12] 杨国良. 国际 VoIP 流量特征分析[J]. 电信科学, 2007,57(6): 7-16.

[13] CAMPELL W M, SINGER E, TORRES-CARRASQUILLO P A, *et al*. Language recognition with support vector machines[C]//Proc of IEEE Odyssey: The Speaker and Language Recognition Workshop. 2004: 41-44.

[14] 邓乃杨, 田英杰. 支持向量机: 理论、算法与拓展[M]. 北京: 科学出版社, 2009: 32-94.

[15] WITTEN I H, FRANK E. Data mining: practical machine learning tools and techniques[M]. 2nd ed. San Francisco: Elsevier Inc, 2005.