硕士学位论文

基于机器学习的 VoIP 流量识别技术研究

论文作者：张峰

指导教师：肖诗松 副教授

学科专业：计算机系统结构

研究方向：网络管理

华中师范大学计算机学院

2013 年 5 月

硕士学位论文
MASTER'S THESIS

The Study of VoIP Traffic Identification Technology Based on Machine Learning

A Thesis

Submitted in Partial Fulfillment of the Requirement

For the M.S. Degree in Computer Science

By

Zhang Feng

Postgraduate Program

School of Computer

Central China Normal University

Supervisor: Xiao Shisong

Academic Title: Associate Proffessor

Signature

Approved

May. 2013

华中师范大学学位论文原创性声明和使用授权说明

原创性声明

本人郑重声明：所呈交的学位论文，是本人在导师指导下，独立进行研究工作所取得的研究成果。除文中已经标明引用的内容外，本论文不包含任何其他个人或集体已经发表或撰写过的研究成果。对本文的研究做出贡献的个人和集体，均已在文中以明确方式标明。本声明的法律结果由本人承担。

作者签名：张峰　　日期：2013 年 6 月 4 日

学位论文版权使用授权书

学位论文作者完全了解华中师范大学有关保留、使用学位论文的规定，即：研究生在校攻读学位期间论文工作的知识产权单位属华中师范大学。学校有权保留并向国家有关部门或机构送交论文的复印件和电子版，允许学位论文被查阅和借阅；学校可以公布学位论文的全部或部分内容，可以允许采用影印、缩印或其它复制手段保存、汇编学位论文。（保密的学位论文在解密后遵守此规定）

保密论文注释：本学位论文属于保密，在____年解密后适用本授权书。

非保密论文注释：本学位论文不属于保密范围，适用本授权书。

作者签名：张峰　　导师签名：

日期：2013年6月4日　　日期：2013年6月5日

本人已经认真阅读“CALIS 高校学位论文全文数据库发布章程”，同意将本人的学位论文提交“CALIS 高校学位论文全文数据库”中全文发布，并可按“章程”中的规定享受相关权益。同意论文提交后滞后：□半年；□一年；□二年发布。

作者签名：张峰　　导师签名：

日期：2013年6月4日　　日期：2013年6月5日

摘 要

VoIP 是建立在互联网上的通信模型，传递用户通信的语音、视频及文本信息。VoIP 有别于传统电话，也更优于传统电话。首先，提供多样性的服务。传统电话仅局限于提供语音服务，VoIP 除提供语音服务外，还提供视频和文本传输服务。以及为多用户提供视频会议服务。其次，提供比传统电话更方便快捷的服务。随着互联网的普及，无线服务覆盖面积的快速增长，用户能够不受时间空间的限制，使用 VoIP 服务。更重要的是相较传统电话服务，VoIP 业务更加便宜。最后，可扩展性更强。使用不同 VoIP 协议标准开发的应用程序之间可以相互通信，并且能够与传统电话通信。这些优势推动着 VoIP 的迅速发展，用户增多的同时，伴随着管理方面问题的出现，而有效管理 VoIP 的基础是通过流量检测，区分正常流量和非法流量。

本文首先简要介绍了 VoIP 及其相关协议标准和技术，阐述流量识别的重要性，分析当前常用的几种 VoIP 流量识别方法。常用的方法主要是基于分析主机和流量行为特征分析，还有一些是通过分析 VoIP 传输时使用的协议实现。在此基础上，本文提出基于机器学习的流量识别方法。该方法对已知流媒体流量的特征进行提取，在 246 个备选特征集合中，选择用于贝叶斯网络训练的核心特征向量。然后利用贝叶斯网络训练原始流量样本，结合特征向量达到 VoIP 流量识别的目的。即分离出 VoIP 及其他软件等正常流量，同时区分非法流量，例如受到 DoS 攻击的流量。通过实验数据分析，并与其他流量识别方法对比，该方法精度、查准率和查全率等指标都较高，能够准确识别 VoIP 流量。最后分析输入特征个数与 VoIP 流量识别方法精度之间的关系，以达到在最少输入的情况下获得最大精度，同时降低识别方法运行时间复杂度和空间复杂度。

随着云计算技术及其他通信技术的快速发展，网络产生的流量呈指数增长，同时实时监控的需求在不断提高。本文提出的方法为海量数据和实时监控也提供了值得借鉴的理论和模型。

关键词：VoIP；流量识别；机器学习；贝叶斯网络

Abstract

VoIP is a communication model based on the Internet, which can transfer voice, video and text information of user communication. VoIP is different from traditional phones, and it is better than traditional phone. Firstly, it provides a diversity of services. The traditional telephone is limited to the provision of voice services. In addition to the provision of voice services, VoIP also provides video and text transmission services. As well as providing video conferencing services for multi-user. Secondly, it provides more convenient and efficient service than traditional phone. With the popularity of the Internet, the rapid growth of wireless services coverage area, users will be free from the restrictions of time and space to use the VoIP services. What's more, VoIP service is much cheaper compared to traditional telephone service. Finally, it's stronger in expandability. Users can communicate with each other and with the traditional telephone communication during different VoIP protocol standard developed applications. These advantages promote the rapid development of VoIP. With the growth of users and the emergence of management issues, the basic of effective management of VoIP is detection and identification of traffic and distinguishing normal traffic from illegal traffic.

This paper briefly describes the protocol standards and technology of VoIP and elaborates on the importance of traffic identification. Through analyzing the current method of several VoIP traffic identification, the methods are mainly based on the analyzing the characteristics of the host and flow behavior or the analyzing transport protocol of VoIP. On this basis, this paper proposes a traffic identification method based on machine learning. The method extracts known streaming flow characterized in 246 alternative feature set, and selects the core feature vector for the Bayesian network training. After that, to achieve the purpose of VoIP traffic identification by making use of Bayesian network training the raw traffic samples, and combining with feature vector. It is to get the normal traffic of VoIP and other software, as well distinguishing the illegal traffic such as DoS attack traffic. Through experimental data analysis and comparison with other traffic identification method, it shows that the indicators such as accuracy, precision and recall is perfect, which means that the method is able to identify the VoIP traffic accurately. Finally, to analyze the relationship between the number of input feature

VoIP traffic identification method accuracy in order to achieve the minimum input for maximum accuracy, while reducing the identification method run time complexity and space complexity.

With the rapid development of cloud computing and other communication technologies, the network traffic generated by exponential, as well the real-time monitoring was in necessary. The proposed method provides good example for big data and real-time monitoring.

Key words: VoIP; Traffic identification; Machine learning; Bayesian network

目 录

摘 要 ..... **I**
ABSTRACT ..... **II**
第一章 绪论 **1**
1.1 研究背景 1
1.2 研究现状 1
1.3 研究目的及意义 2
1.4 本文主要研究内容和组织结构 3
1.5 本章小结 3
第二章 技术背景 **4**
2.1VoIP 简介 4
2.2VoIP 相关协议 5
2.2.1H.323 协议 5
2.2.2SIP 协议 6
2.2.3RTP/RTCP 协议 8
2.2.4Skype 协议 9
2.3 机器学习 11
2.3.1 隐含马尔可夫模型 12
2.3.2 人工神经网络 13
2.3.3 朴素贝叶斯 15
2.3.4 遗传算法 16
2.4VoIP 流量识别方法 18
2.4.1 基于主机行为特征的流量识别方法 18
2.4.2 基于 SIP 和 H.323 协议的流量识别方法 18
2.4.3 基于流量特征提取的识别方法 19
2.5 本章小结 19
第三章 基于机器学习的 VOIP 流量识别 **21**
3.1 主机行为特征 21
3.2 流量行为特征 22
3.3 机器学习流量识别 23

3.3.1 流量特征选择.....24
3.3.2 贝叶斯网络.....25
3.3.3 贝叶斯网络训练.....26
3.4 本章小结.....29
第四章 实验方法及结果分析.....30
4.1 实验环境.....30
4.2 实验数据集.....30
4.3 分类方法评价指标.....32
4.4 实验结果分析.....32
4.4.1 分类检测结果.....32
4.4.2 不同机器学习算法的比较.....34
4.4.3 特征参数数量与精度之间的关系.....35
4.5 实验总结.....36
4.6 本章小结.....37
第五章 总结和展望.....38
5.1 全文总结.....38
5.2 技术展望.....38
参考文献.....39
在校期间发表的论文、科研成果等.....42
致谢.....43

第一章 绪论

1.1 研究背景

相较于传统的电话通信服务，网络语音电话 VoIP[1]在过去几年得到迅猛的发展。VoIP（Voice over Internet Protocol）是一种以 IP 电话为主，并提供相应增值附加服务的新型语音通信技术[2][3]。VoIP 将用户语音通信的模拟信号转化为能够在互联网上传输的数字信号，实现实时传递[4]。其最大的优势在于充分利用互联网的全球遍及性，同时随着网络质量的不断提高，提供比传统语音通信业务更方便安全高效的服务。另外，无线通信技术的发展及不断改进，使得用户不受时间空间位置的限制可以自由接入互联网。同时互联网提供互联网接入功能的智能手机和移动计算设备的层出不穷，为 VoIP 服务提供更好的硬件支撑。VoIP 除了比传统电话更便宜外，也提供其他的功能服务，例如视频会议。

互联网给用户提供极大方便的同时，其开放性也给用户平添诸多隐患。现实生活中，通过窃取通信软件发出的信息或者直接窃取服务提供商的数据，从而获得用户隐私数据的案例屡见不鲜。VoIP 系统所承受的风险不仅存在于公共数据网络，而且与 VoIP 所采用的相关协议不无关系。例如，潜在的恶意用户有可能能够捕获 VoIP 数据包，重构相关数据以实现数字窃听[5]。VoIP 服务在用户之间传递语音和视频数据，需要占用较多的网络带宽，增加网络运营商的成本。当网络中存在 VoIP 攻击，造成 VoIP 包泛滥时，会导致网络瘫痪[6]。

为了确保互联网络和移动网络的正常运行，避免资源浪费，达到资源的高效合理使用，需要建立流量检测机制。流量检测机制通过分析在网络中传输的流量数据，通过识别算法，区分正常流量和非法流量数据包。对于存在的威胁，能够给出有效解决方案，及时消除攻击源，确保网络畅通。在这方面，国内外很多专家学者都进行了大量的研究，提出多种检测算法，形成合理的解决方案。

1.2 研究现状

VoIP 的发展带来机遇的同时给网络正常运行带来挑战，比如网络的安全和网络流量控制。传统对文本数据流量分类识别的方法，处理的数据限于文本信息，对于 VoIP 大数据流量的检测不能胜任。认识到这一点，大量的科研机构和个人在 VoIP

流量识别方面进行研究，取得了很多可观的科研成果，并将这些成果应用到实际的网络环境中，为网络安全做出不可磨灭的贡献。

国外是最早研究流量识别技术的，而且提出了很多理论模型。Yongmin Choi 利用软件签名技术分类互联网网络流量，大幅提高流量识别的精确度[7]。Hun-Jeong Kang 等人基于多媒体服务协议分析提出流量识别算法[8]。Tom Auld、Andrew W.Moore 和 Stephen F.Gull 利用贝叶斯技术分类网络流量，推广人工智能技术在网络安全领域的应用[9]。Ping Du 等人根据数据报文长度的分布情况，检测网络中存在的 DoS 攻击，提高网络安全[10]。Ehlert S 等人分析 Skype 软件登录过程，并提取应用层协议特征字和端口号，提出根据 Skype 报文长度和顺序来识别 Skype 的方法[11]。Bijan Raahemi 等人通过发掘 IP 数据流的特征，使用快速决策树识别 P2P 流量[12]。

在国内，也有很多学者对流量识别的研究做出贡献。杨国良等人对 VoIP 流量分布和协议分布进行了大量研究[13]。王振华等人主要利用特征提取的方法对 Skype 产生的流量进行检测[14]。Li Jun 等人从网络分层出发，主要收集传输层数据包，使用遗传算法进行流量分析识别[15]。王蕊等人的出发点主要是 Skype 的流量特征，通过对流量的特征值提取进行分析[16]。Li Bing 等人基于主机和流量的行为进行分析，实现对网络流量特别是 Skype 流量的检测[17]。

国内外对 VoIP 的研究主要集中在具有代表性的 Skype 通信软件，但是现在的网络情况越来越复杂。比如，国内的 QQ 软件也提供语音通信服务，所使用的协议算法和 Skype 软件有所不同。传统基于 Skype 通信软件所作的研究，对 QQ 通信软件使用的有效性有待进一步验证。另外随着软件种类的增多，产生的流量包的类型也在增多，对流量识别算法的精确度、查准和查全率方面都有影响，达不到网络检测的要求。

1.3 研究目的及意义

VoIP 应用软件的推广，软件类型的增多，用户数据的膨胀，这一系列的因素都在不断影响网络安全。一方面，网络协议本身的漏洞为非法用户提供入口，威胁用户数据安全。另一方面，用户增多数据流量增大，DoS 泛洪攻击出现的可能性增大，网络面临瘫痪的危险。因此分析网络流量，检测产生非法流量的根源，消除网络威胁是重中之重，也是网络正常工作的根本。

流量识别的实现主要有两方面的作用：第一，确保网络安全。及时发现存在威胁的流量，由系统或人为地中断非法流量的传播，保证网络安全。第二，方便网络管理。流量识别能够标识产生流量的网络软件，对于产生攻击流量的软件，管理人

员可以对其屏蔽。对于带宽有限的网络，这是非常重要的。

1.4 本文主要研究内容和组织结构

本文主要研究内容主要包括，第一，总结传统流量识别算法；第二，发掘可以改进的切入点，提出改进的基于机器学习的流量识别方法。首先，对 VoIP 相关概念进行阐述，大致了解什么是 VoIP。其次，分析 VoIP 使用到的协议。最后分析当前存在的流量识别算法，比较算法之间的优缺点。提出改进的基于机器学习的流量识别算法并检验有效性和准确性。

本文结构安排如下：

第一章简要介绍 VoIP 流量识别的背景，研究现状和研究目的及意义以及本文的主要内容。

第二章简要介绍与 VoIP 相关的几个网络协议，并对机器学习的两种方法进行分析，最后分析当前常用的几种流量识别方法的优缺点。

第三章简要介绍传统流量识别方法，分析这些方法应用的局限性和存在的不足，并在此基础上提出贝叶斯网络分类方法，给出评价指标。

第四章首先对实验数据和实验结果进行分析，突出其有效性。和其他分类方法比对，突出其准确度。

第五章总结本文研究内容，并提出文中没有涉及的可以进一步研究的问题。

1.5 本章小结

本章主要介绍了 VoIP 流量识别研究的背景，及当前国内外对流量识别方法的研究的现状。阐明本文研究的目的及意义，描述本文的组织结构。

第二章 技术背景

2.1VoIP 简介

VoIP（Voice over Internet Protocol）对传统电话模式具有颠覆性的改革，并且方便廉价，因此在很短时间内得以推广应用。VoIP 为用户提供一系列功能，其中包括基本的语音通话，还包括视频会议、短信消息及传真等多媒体服务。由于 VoIP 是通过网络传输用户之间通信的数据和信息，因此这种技术有时也被人们称作网络电话或 IP 电话。

VoIP 和其他网络软件一样，在数据交换方面，都遵循相关的网络协议，本文后面将简要介绍几种常用的语音通信协议。在网络中传输的数据，首先会被分割成段，每一段被称作数据包[18]。这些数据包从通信的源端发向目的端。每个数据包包含相应的报文头部，在报文头部中含有目的端的信息，例如目的端的地址等。数据包依据头部信息以及网络带宽等信息在网络中传输，宽带信息会影响数据包传输的路径选择。典型的数据交换网络是互联网，同时无线网络和移动通信网络也在不断发展。虽然 VoIP 相对传统应用传输的是语音信息，产生的瞬间流量较大，对网络设备的要求相对较高。但是现在的网络基础设施已广泛存在，这些设备为 VoIP 的运行提供充足的硬件平台，这也就使 VoIP 代替传统模拟电话服务成为可能。因为接入网络变得越来越简单，即使使用 VoIP 不是免费的，相对传统电话来说，它也是很便宜的。

传统模拟电话运行于公共交换电话网络。公共交换电话网络是一种基于虚拟电路交换实现的网络。电路交换网络实质上在两个用户之间建立专用连接，以此来确保提供高带宽低延迟的高质量服务。高质量的服务使得通话用户能方便地携带移动电话，同时不会对通话质量产生影响，因为网络延迟对通话造成的影响将是用户所不能接受的。在这方面 VoIP 存在不足，它并不能确保提供高质量的数据交换服务。因为 VoIP 工作在互联网上，互联网的数据传输会产生拥塞，导致数据不能及时传递给目标用户。这也和互联网的设计有关，因为互联网不是设计用来传输实时数据的。

VoIP 数据传输时，首先需要将用户的语音信息通过编码器转化成数字信号，而且编码器会折中语音质量和带宽效率，确保传输成功率。当语音信号转化为数字信号，将作为一系列数据包在网络上传输。传输过程中会受到诸多因素影响，比如潜

在的网络攻击、数据包的丢失、网络带宽、数据包乱序以及最常遇到的网络拥塞。这些因素在数据交换网络中都是人为难以控制的，而这些因素对服务质量会产生极大的影响。McKnight 等人在 2001 表示当 VoIP 数据包开始传输时，网络就变得不可控制。因此 VoIP 在语音通话领域显然不是一个强劲的竞争者。这同时也是 VoIP 发展的主要障碍，但是研究者仍在努力使 VoIP 适应现在的网络。

经过多年的发展，现存有很多 VoIP 标准协议供用户公开使用，这些标准定义了两个或更多用户的语音通话。在这些标准协议中，最重要的 H.323 和 SIP，它们都是开放标准，使得在用户使用不同的 VoIP 产品也可以交流。

2.2VoIP 相关协议

网络协议是计算机互联网正常运行的基础，规定数据交换的规则。最常用的网络协议有网络层的 IP 协议、ARP 协议、RARP 协议等，传输层的 TCP 协议、UPD 协议等，应用层的 FTP 协议、HTTP 协议等。每种类型的软件也可以自定义相关协议，这里不作介绍。与 VoIP 相关的协议主要有 RTP/RTCP 协议、H.323 协议、SIP 协议和 Skype 协议等，如图 2-1 所示。

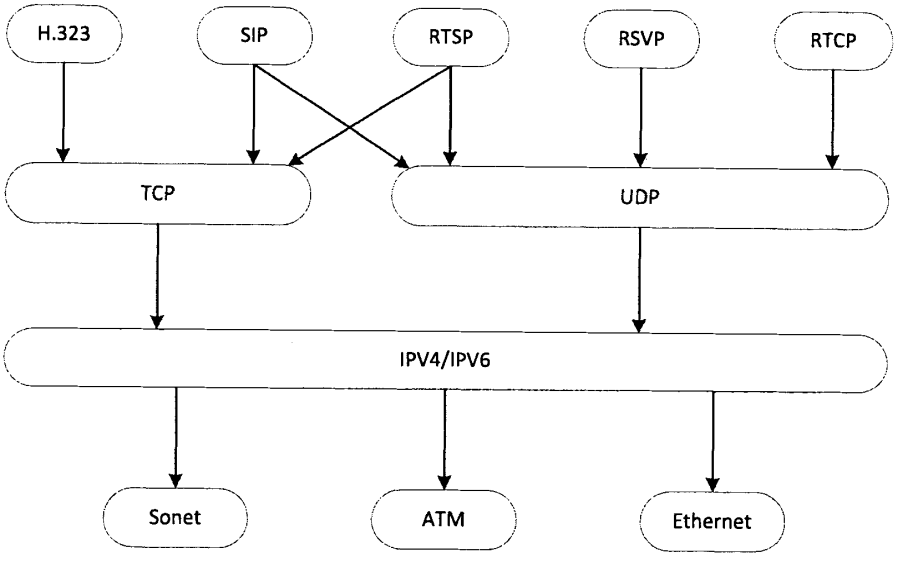


图 2-1VoIP 协议

2.2.1H.323 协议

H.323 是 ITU-T 于 1996 年提出的推荐协议标准，至 2009 年已更新维护至第 7 版。H.323 主要用于描述分组传输网络通信的实体，这些通信实体并不能保证稳定的服务质量。实体可以提供的服务包括实时语音、视频和数据通信。语音服务是使

用该协议的终端必须提供的服务，而视频和数据通信服务则是可选的[19]。因此使用H.323 协议开发终端可以相互通信，支持点对点的通信，同时支持复杂网络拓扑的多用户通信。

H.323 协议定义了使用该协议的终端进行多媒体通信的系统技术要求，系统的底层是不能提供稳定服务质量的分组传输网络。同时还定义了 H.323 系统的核心组成部分，包括终端、网关、网守[20]、多点控制器、多点处理器和多点控制单元等[19]。其中网关是终端和其他 H 系列终端通信的接口，例如通用交换电话网络、综合业务数字网语音终端，通用交换电话网络、综合业务数字网数据终端等，如图 2-2 所示。多点控制器、多点处理器和多点控制为多点会议通信提供支持。

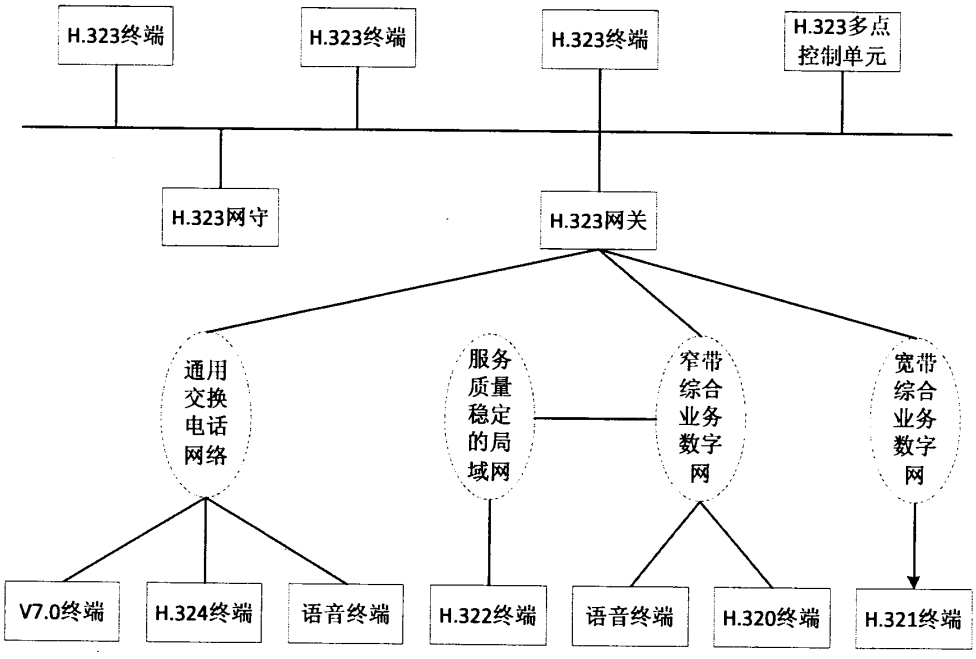


图 2-2 ITU-T H.323 终端连通性

2.2.2SIP 协议

会话发起协议第一个标准版本是由 IETF 在 1999 年发布的，主要作为多媒体通信协议使用，处理通信会话客户端和服务器间的控制报文，比如 VoIP 通话。由于 SIP 在会话期间是独立工作的，因此不仅可以用于文本信息，同时也可以用于视频会议。最新的 SIP 版本是 2.0，于 2002 年发布。

SIP 处理 VoIP 通话必需的控制报文，这些报文通常用来确定与通话用户相关的重要信息，包括用户位置、可用性以及性能参数[21]。SIP 同时也用来建立会话并管理会话。报文主要有两种类型，即请求报文和响应报文。SIP 类似于 HTTP（Hypertext

Transfer Protocol）协议，同样遵循请求/响应通信模型。对于每一个由 SIP 客户端发出的请求，必须有一个或多个对应的响应返回，这样就形成了一个完整的通信过程。这种由客户和服务器参与的通信就是通常意义上的会话。

一般消息 = 起始行
消息包头
CRLF
[消息正文]
起始行 = 请求行/状态行

图 2-3 SIP 协议消息正文

图 2-3 所示为 SIP 协议消息正文及类型的结构，它的定义基于网络信息结构，使用 UTF-8 字符集。SIP 消息包括客户端连接服务器端的请求，以及服务器端返回给客户端的响应。由于不同的系统在字符集和语法上会有所不同，因此会造成消息处理上的不便，因此对消息格式的定义统一遵守 RFC2822 的约定，不论是请求消息还是应答消息。起始行是请求/应答消息都必须包含的。消息包头的数量是一个或多个。消息正文是可选的。结构中的 CRLF 表示回车。起始行、每个包头行、空行都必须以回车换行结束。即使消息只包含前面的两个部分而没有正文，也必须以一个空行结束。

SIP 消息结构中，开始行后会跟随一个或多个头。SIP 头的定义和 HTTP 头的定义在语法和语意上比较类似，遵循 H4.2 消息头语法规则。除些之外，SIP 头定义也遵循多行扩展头定义规则。SIP 头是由一个域名加上冒号（":"）和域值组成。

$$field-name: field-value$$

在消息头的定义中，允许冒号的左右有一到多个空白，但是为了方便维护，在实现时，域名和冒号之间一般不放置空格，冒号和域值之间可以适当添加一或两个空格以控制格式方便阅读。对于多行扩展头的定义，只要定义每一个附加行以至少一个空格或者 TAB 开头即可。

请求消息的起始行由方法、请求类型统一资源定位符和 SIP 版本号组成。方法表示的是处理服务器请求的主要功能，是请求消息自身携带的。典型的方法是 INVITE 和 BYE。RFC3261 预定义了六个主要方法：REGISTER、INVITE、ACK、CANCEL、BYE 和 OPTIONS，通过 SIP 扩展可以增加更多的功能。请求统一资源定位符标识请求发送的目的方。下面是一个 SIP 统一资源定位符的示例：

$$sip:user@192.168.1.1:5060$$

统一资源定位符通常以 sip 开头，@符号前的 user 标识主机上的资源，本例中的主机是指 192.168.1.1，主机是 IP 地址或者完整的域名。5060 代表主机的端口号，并且 5060 是 SIP 使用的默认端口号。

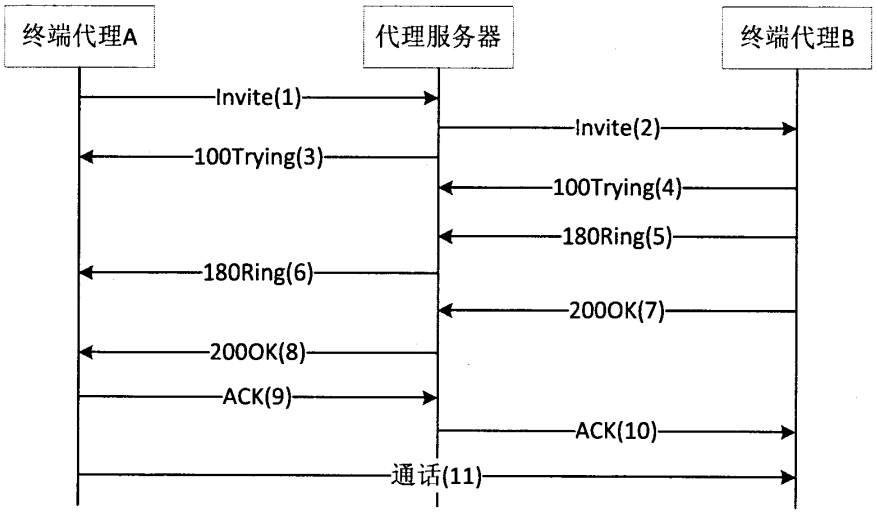


图 2-4SIP 基本呼叫流程图

SIP 基本呼叫建立过程如图 2-4 所示。首先用户发起呼叫，终端代理 A 向隶属的代理服务器发送 Invite 请求。代理服务通过用户的认证后，检查请求信息 Invite 是否合法，对不合法的请求，返回错误信息。提取合法请求的目的地址域，并将相应的请求转发至被叫终端代理 B。100Trying 是终端代理 A 和终端代理 B 经代理服务器发送处理的应答消息。180Ring 信息是由被叫用户终端代理 B 向代理服务器发送，再转发给发起呼叫的主叫用户终端代理 A。此过程结束后，如果被叫用户应答此次呼叫，则终端代理 B 会向代理服务器发送表示连接成功的应答信息 200OK，再由代理服务器转发给终端代理 A。终端代理 A 收到 200OK 的应答信息后，会经由代理服务器向终端代理 B 发送确认信息 ACK。终端代理 B 收到确认信息 ACK 表示此次呼叫通信成功建立。

2.2.3RTP/RTCP 协议

RTP 是实时传输协议 Real-Time Transport Protocol 的缩写，RTCP 是实时传输控制协议 Real-Time Transport Control Protocol 的缩写[22]。流媒体通信的显著特点是传输数据量大，使用传统 TCP 协议，在网络不理想出现丢包时会激活重传机制和拥塞控制机制，虽然能够提高服务质量，但也会造成网络资源浪费。因此需要专门满足流媒体实时通信的协议，RTP 就是这种协议，它工作于 UDP(User Datagram Protocol)之上。由于 UDP 不能像 TCP 那样能够提供安全可靠的服务，因此使用 RTCP 来保

证实时数据传输的效率的质量。除了 RTCP 协议工作在 TCP 层上，实时流协议（Real-Time Streaming Protocol）也工作在 TCP 层上，起着控制作用，如图 2-5 所示。

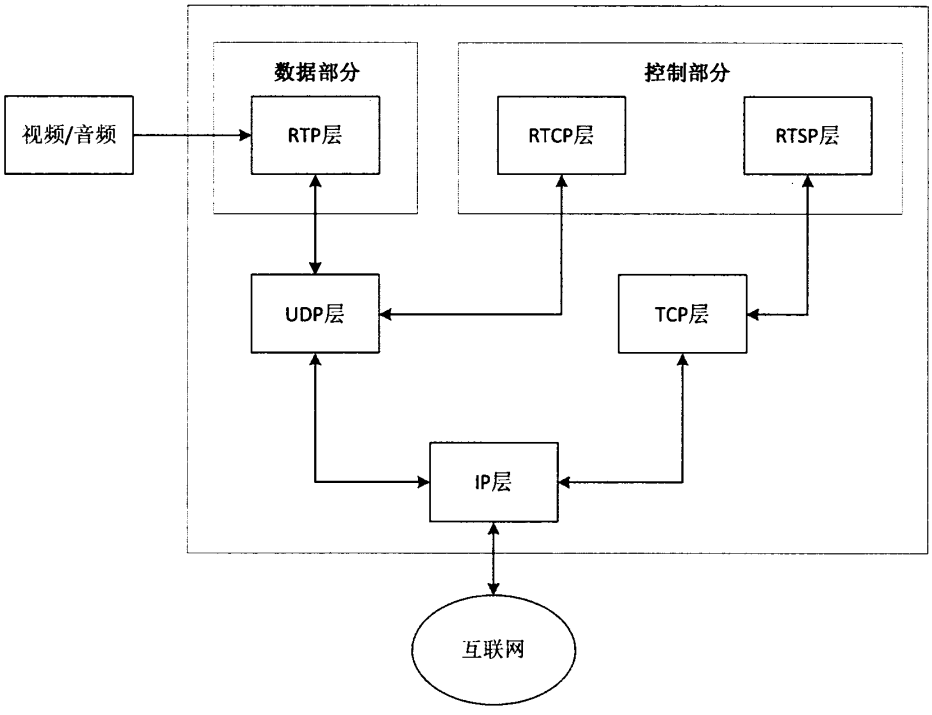


图 2-5 RTP、RTCP 和 RTSP 协议

流媒体传输启动时，服务器会收到 RTSP 请求并发生 RTSP 请求对象。服务器使用 RTSP 协议应答信息将请求的内容以流会话的形式传递，内容包括媒体类型、所使用的编码格式和流数据。一个流会话可以只包含一个流，也可以是由多个流组成，作用主要用于提供时间信息和实现流同步。RTP 本身是工作在 UDP 层上，不能提供可靠的顺序传输服务，流量控制和拥塞控制主要由 RTCP 协议完成。在 RTP 会话期间，连接者时刻监视网络使用情况记录成 RTCP 包，并周期性发送通知发送方。RTCP 包信息包括已发送数据包的数据，网络丢包情况等。由此发送方可以根据这些信息更好地利用网络。服务器也可以根据这些数据决定发送数据的频率和每次发送的数据，即实现动态改变数据传输速率。通过 RTP 和 RTCP 协议的配合使用，可以有效地反馈网络利用率，提高传输成功率。

2.2.4Skype 协议

Skype[3]是 KaZaa 开发的 P2P 通信客户端软件，可以为用户提供语音、视频以及文本信息传输服务，与 MSN、Yahoo 即时通信及 QQ 等软件很相似，但是 Skype 使用的协议与其他即时通信软件是不同的。Skype 网络拓扑结构如图 2-6 所示，是

一个覆盖 P2P 网络，由普通主机和超级节点组成。普通主机是一个 Skype 应用程序，用于处理语音通话和文本信息传输。超级节点在 Skype 网络上是普通用户的终端。节点成为超级节点的条件配置公共 IP 地址，CPU、内存和网络带宽充足。普通主机必须连接到超级节点，并在 Sky 登陆服务器注册，才能成功登陆。虽然 Skype 登陆服务器并不是 Skype 网络中的节点，但是它在 Skype 完成通信的整个过程中是不可或缺的。首先用户的注册信息（包括最主要的用户名和密码）都存储在 Skype 登陆服务器，用户登陆认证功能也由 Skype 登陆服务器完成。其次，用户注册时，登陆服务器检测是否已经被注册，只有没被注册的用户名才能注册登陆。

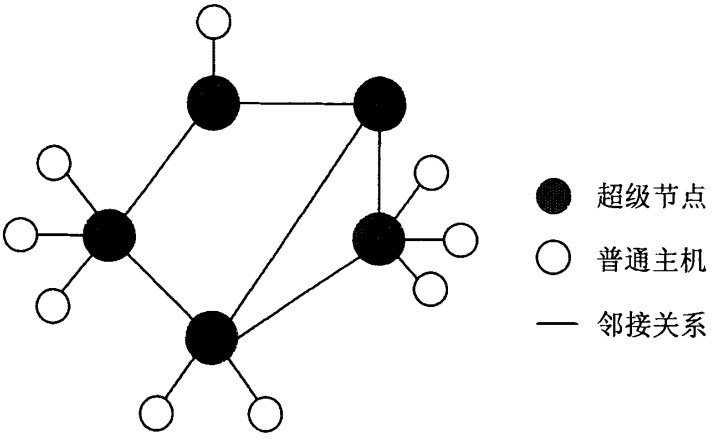


图 2-6Skype 网络拓扑结构

由于 Skype 网络是一个覆盖网络，因此每个节点必须独立建立和维护一个可达节点表，这个表被称为主机列表，记录超级节点的 IP 地址和端口号，保存在操作系统注册表中。Skype 实现的全局索引技术可以利用注册表中的信息找到过去 72 小时内登陆 Skype 网络的用户，缩短通信建立时间。

Skype 使用宽带编码技术，确保在 32kb/s 的宽带环境中也能提供理想的通话服务。Skype 使用 TCP 协议在用户之间传输控制信号，同时使用 TCP 和 UDP 协议传输媒体流量，控制信号和媒体流量的发送使用不同的端口号。

Skype 软件主机的主要组成部分包括：端口号、主机列表、编码器、好友列表、加密技术、网络地址转换和防火墙。Skype 与其他软件不同，没有默认的端口号，在软件安装时会随时选择一个端口作为其服务监听端口，同时会监控 80 和 433 端口。使用过程中，也可以在连接属性对话框中设置端口号。主机列表就是前面提到的保存在操作系统注册表中的超级节点 IP 地址和端口号，对应的键为 HKEY_CURRENT_USER/SOFTWARE/SKYPE/PHONE/LIB/CONNECTION/HOSTC

ACHE，且最多只能保存 200 条记录。编码器是 Skype 的核心模块，负责完成用户传输的语音、视频及文本编码工作，采用 iLBC、iSAC 及第三方编码器。好友列表被保存在本地操作系统注册表中，而且经过数字签名和加密，确保安全性。如果用户在其他主机上登陆 Skype 网络，那么该用户必须重新建立好友列表。加密技术是确保用户通信安全的关键。Skype 使用 256 位的 AES 加密算法，总计有11×10^{77}种可能密钥。为了更有效地对通信数据进行加密，Skype 使用 1536 至 2048 位的 RSA 来协商对称 AES 密钥。Skype 绑定的网络地址转换和防火墙信息也是保存在本地操作系统注册表中，而且是定期更新的。

Skype 相比于其他即时通信软件的技术优势在于：一、Skype 采用 P2P 技术，P2P 相当于将每个客户端作为服务器，因此可以无限扩展，不会因扩展带来搜索时间的延长。在数据传输方面采用智能路由算法，保证传输高可达率的同时保证较高的通话质量。二、Skype 采用 AES 加密算法，动态加密每个呼叫和即时信息，即使在传输过程中信息被截取，被破解的可能性几乎为零。三、由于 Skype 运行端口号不是固定的，因此可以不受防火墙等工具的影响正常运行。四、由于 Skype 开源主功能模块，因此开发者可以从官方网站下载类库，开发不同平台的通信软件，其核心依然是 Skype。五、传输大型文件效率较高。

2.3 机器学习

机器学习是基于建立显式或隐式模型对所分析的实体进行分类，是近年来计算机科学研究的热门领域。这些模型的典型特征是需要标记数据训练行为模型，这个训练过程占用较多的系统资源。计算机完成特定功能的前提是具有相应的算法，例如常用的排序算法、文件压缩算法等。有些功能使用计算机来处理能极大提高工作效率，但是这些功能的实现很难准备地用算法描述，例如光学字符识别、语音识别和面部识别功能。就面部识别而言，使用传感器能够获取人脸图像，但是我们更多是希望利用图像对比辨认特定个体。虽然个体实处的环境不同，衣着外表也有所变化，但是我们依然可以轻松辨认不同的人，困难的是我们很难用计算机只可以理解的算法来描述辨认过程。因此就需要使用机器学习通过自学习功能完成诸如此类的任务。

机器学习的一个重要应用是分类。例如一个银行想要开发一个系统用来区别真币和假币，输入信息包括纸币的颜色、大小和尺寸。机器学习通过对一组真币数据的分析，训练辨识能力，最终能够纸币的真假。这种情形也叫监督式学习，发生在

有一系统输入和输出的情况下，目的是发掘输入和输出之间的关系。

机器学习方法主要有两种，即隐马尔可夫模型和人工神经网络。

2.3.1 隐含马尔可夫模型

隐含马尔可夫模型（HMM，Hidden Markov Model）是遵循马尔可夫性质的系统统计模型，因此也被称为马尔可夫过程[23]。马尔可夫性质是指，一个随机过程在已知的现在状态及所有过去状态的情况下，未来状态的条件概率分布仅依赖当前的状态[24]。隐含马尔可夫模型是由一系列的状态以及这些状态之间的转变、可观察的结果和可能和输出组成。隐含马尔可夫模型状态是不明显的，学习过程决定状态转换概率。这个模型可以用于确定特定系统基于观察输出序列的最可能的状态转换序列。

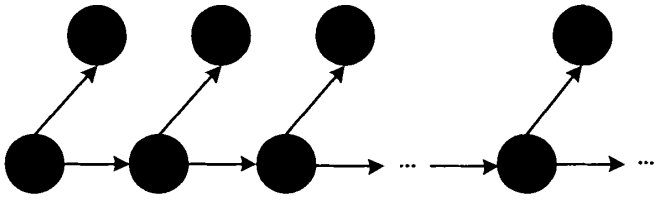


图 2-7 隐含马尔可夫模型

如图 2-7 所示，是一个典型的隐含马尔可夫模型。给定序列$s_1,s_2,s_3,\cdots$产生输出$o_1,o_2,o_3,\cdots$。随机变量s_t表示当前时刻t的隐含状态，随机变量o_t表示当前时刻t的观察状态，而且o_t仅和当前时刻t的状态s_t相关联，被称为独立输出假设。箭头表示各状态之间的条件依赖关系，即随机过程中状态s_t的概率分布，仅与邻接的前一个状态s_{t-1}相关。这种依赖关系的假设未必适合所有的情况，对于要求不高的分类情况，这种假设是适用的。

隐含马尔可夫模型的隐含变量的状态空间是彼此独立的，观察输出结果可以是彼此独立也可以是连续的。隐含马尔可夫模型定义转换概率和发散概率，发散概率也可以说是输出概率。转换概率控制隐含状态转换路径选择的可能性。假设隐含状态空间包含N个可能值，时刻t的状态可能是N个状态中的任意一个，从当时时刻转换到$t+1$时刻的状态（N个状态中的任意一个）总共有N^2种转换路径。需要注意的是从任意给定的状态转换的转换概率集合的总和必须为 1。因此$N\times N$的转换概率矩阵是一个马尔可夫链，因为已知其他转换概率可以确定任一转换概率，总共

有 $N(N-1)$ 个转换参数。对于 N 个可能状态中的任意一个，发散概率集合用于描述特定时刻隐含变量观察结果的分布情况。集合的大小依赖观察变量的性质。例如，如果观察变量是相互独立的 M 个可能值且服从类别分布，将有 $M-1$ 个独立的变量，所有的隐含状态总共有 $N(M-1)$ 个发散参数。如果观察变量是一个 M 维的向量，服从任意多元高斯分布，则有 M 个表示平均数的参数和 $M(M+1)/2$ 个表示协方差矩阵的参数，共计 $N(M+M(M+1)/2)=NM(M+3)/2=O(NM^2)$ 个发散参数。由此可见，除非 M 很小，否则带来的计算量将很大，因此假设观察变量是相互独立的，将极大降低计算量。

2.3.2 人工神经网络

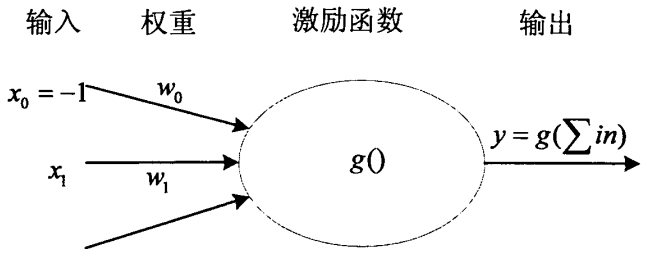


图 2-8 神经网络节点

除了隐含马尔可夫模型，机器学习的另外一种方法是人工神经网络[25]。人工神经网络的设计主要用来模拟人类大脑的工作过程。神经网络由相互连接的神经（或称为节点）组成，与生物学中的神经元相对应。神经网络每个结点的结构如图 2-8 所示。x 表示节点的输入，可以是外部数据源，例如传感器产生的数据，也可以是其他结点的输出。加权值被称为权重，是每两个节点间的连接，影响输出结果。每个节点有一个固定的输入 $x_0=-1$，连接处的权重为 w_0，设置节点的生效临界点。输出的结果是所有的输入乘以各自相关的权重，相加求和，导入激励函数计算得到。

典型的激励函数是阈值函数（threshold function）和 S 型函数(Sigmoid function)[26]。如果输入的值小于等于 0，阈值函数的结果是 0；如果输入的值是正数，阈值函数结果是 1。S 型函数的定义为 $1/(1+e^{-x})$。S 型函数的优势在于它具有可微性，是训练算法有效的保证。

图 2-9 所示为三层前馈神经网络，分别为输入层，隐藏层和输出层，每一层从

前一层接收输入数据。三层前馈神经网络包含多个隐含层，一般只包含一个输入层和一个输出层[27]。节点 1 和节点 2 表示输入层，节点 3 和节点 4 表示隐藏层，节点 5 表示输出层。输入层接受来自神经网络外部的输入数据，乘以各自的权重后输出给激励函数处理，最终传送给输出结点。神经网络的输出可以用以下公式(2.1)表示：

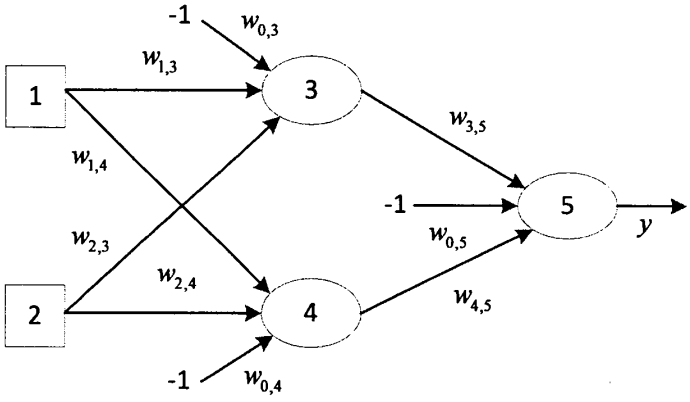


图 2-9 三层前馈神经网络

$$\begin{aligned} y_3 &= g(-w_{0,3} + x_1 w_{1,3} + x_2 w_{2,3}) \\ y_4 &= g(-w_{0,4} + x_1 w_{1,4} + x_2 w_{2,4}) \\ y = y_5 &= g(-w_{0,5} + y_3 w_{3,5} + y_4 w_{4,5}) \end{aligned} \tag{2.1}$$

式中 x_1 和 x_2 对应节点 1 和 2 的输入，相应的 y_i 对应节点 i 的输出。

激励函数表示人工神经网络中单个神经元的输入与输出之间的函数关系，典型的输出是 0 到 1 之间的随机数，在分类应用中，输出有可能是布尔值。当用作分类时，一般以 0.5 作为分隔界限，在 0.5 及以上的属于一类，0.5 以下的归属于另一类。

反向传播算法用来训练多层前馈神经网络。训练集中的每个实例通过神经网络传播计算得到最终的输出结果。将输出结果与预期结果相比较，得到训练实例平方误差的总和。x 表示输入，h 表示网络输出，y 表示输出结果。每个节点的输出误差可由公式(2.2)表示。

$$Err = \frac{1}{2}(y-h)^2 \tag{2.2}$$

其中 $h = g(in) = g(\sum x_i w_i)$。误差通过网络向后传递作为权重的一部分更新计算规则，权重的更新是梯度递减的。输出层给定节点每个权重误差的偏导数的计算方法如公式(2.3)所示：

$$\begin{aligned}\frac{\delta E}{\delta w_i} &= Err \cdot \frac{\delta Err}{\delta w_i} \\ &= Err \cdot \frac{\delta}{\delta w_i}(y - g(\sum_i x_i w_i)) \\ &= -Err \cdot g'(\sum_i x_i w_i) \cdot x_i\end{aligned} \tag{2.3}$$

其中i表示节点的每个输入分支。权重更新的规则如公式(2.4)所示

$$w_i = w_i + \alpha \cdot g'(in) \cdot x_i \tag{2.4}$$

其中 α 表示学习率，控制每次权值改变的幅度。

前隐层节点权重更新规则遵循相同的思路，误差从输出层相应地向后传递至与隐藏节点相关的权重。隐藏层节点权重更新方法如公式(2.5)所示

$$w_{j,i} = w_{j,i} + \alpha \cdot y_j \cdot g'(in) \cdot \sum_i w_{j,i} \cdot Err_i \cdot g'(in_i) \tag{2.5}$$

其中 j 表示隐藏节点，i 表示输出层节点。在节点权重改变之间误差从输出层反向传播，这个过程一直持续到误差传播至输入层或者所有节点的权重都更新为止。

该算法对训练集中的每个实例都执行，遍历一次训练集后，也意味着一轮训练的结束。通常情况下，神经网络会进行指定次数的训练，直到权重维持在稳定的范围内，整个训练过程都会结束。

2.3.3 朴素贝叶斯

朴素贝叶斯方法是基于强贝叶斯独立性假设的简单概率分类器，即假设用于分类的各个特征之间是相互独立的，不存在任何的关联性。贝叶斯分类器每次处理特征集合中的一个单独特征，不管其余的特征。

在多种概率模型中，朴素贝叶斯分类器能够在监督学习下实现高效训练。一些特殊的应用使用极大似然方法计算参数估计。换句话说，在不服从贝叶斯模型或不使用贝叶斯方法时，可以使用朴素贝叶斯模型。虽然朴素贝叶斯的假设显得不是很理想，但是它对一些实际问题的处理仍能收到较好的效果。

朴素贝叶斯网络的一个优点在于，它可以只使用较少的训练数据达到评估分类器所需参数的目的。因为该方法假设参数变量是相互独立的，只需要计算每个分类变量的方差，而不是整个协方差矩阵。

朴素贝叶斯可以定义为：设 $x=\{a_1,a_2,...,a_m\}$ 为待分类项，其中 a 为 x 的一个特

征值。同时存在类别集合$C=\{y_1,y_2,...,y_n\}$。如果有

$$P(y_k|x)=\max\{P(y_1|x),P(y_2|x),...,P(y_n|x)\} \tag{2.6}$$

那么则有$x\in y_k$。换句话说，如果满足公式(2.6)，那么待分类项x属于类别y_k。

$P(y_i|x)$的计算，首先选取已知的待分类项集合（训练样本），统计计算各个类别中每一个特征属性的条件概率估计，如公式(2.7)所示。

$$\begin{aligned}&P(a_1|y_1),P(a_2|y_1),...,P(a_m|y_1);\\&P(a_1|y_2),P(a_2|y_2),...,P(a_m,y_2)\\&\vdots\\&P(a_1|y_n),P(a_2|y_n),...,P(a_m,y_n)\end{aligned} \tag{2.7}$$

由于每个特征都假设是相互独立的，因此$P(y_i|x)$的计算方法如公式(2.8)所示。

$$\begin{aligned}P(y_i|x)&=\frac{P(x|y_i)P(y_i)}{P(x)}\\&=\frac{P(a_1|y_i)P(a_2|y_i)...P(a_m|y_i)}{P(x)}\\&=\frac{P(y_i)\prod_{j=1}^{m}P(a_j|y_i)}{P(x)}\end{aligned} \tag{2.8}$$

2.3.4 遗传算法

遗传算法是一种启发式算法，通过模拟生物进货过程搜索问题的最优解，最初是1975年由美国Michigan大学的J.Holland教授提出的。遗传算法采用自然进化方法寻求问题的最优解，过程包括遗传、变异、选择和交叉，具体流程如图2-10所示。

遗传算法中，优化问题的解一般被称为个体，表示一个变量序列，也叫做染色体。起初，算法随机生成一定量的个体，或者由操作者选择初始生成的个体情况，由此可以提高初始种群的质量。种群的大小取决于问题的规模，一般是生成上百或上千个个体。

确定初始种群后，一部分个体用于产生下一代种群。个体的选择是基于适应来确定的，适应度较优的个体被选择的可能性更大。适应度函数用于计算每个个体的适应度，以此评价一个个体。适应度函数的定义依赖于要解决的问题。例如背包问题，在固定容量的背包里使得放入的物品的价值最大化。此时解决方案的表示可能是一组二进制值，每个二进制值表示一个物品，而0和1则表示是否将此物品背包。

此时的适应度函数就是装入背包的物品价值最大值。但是也有一些问题是很难定义适应度函数的，在这种情况下，通常使用模拟的方法确定个体适应度值。

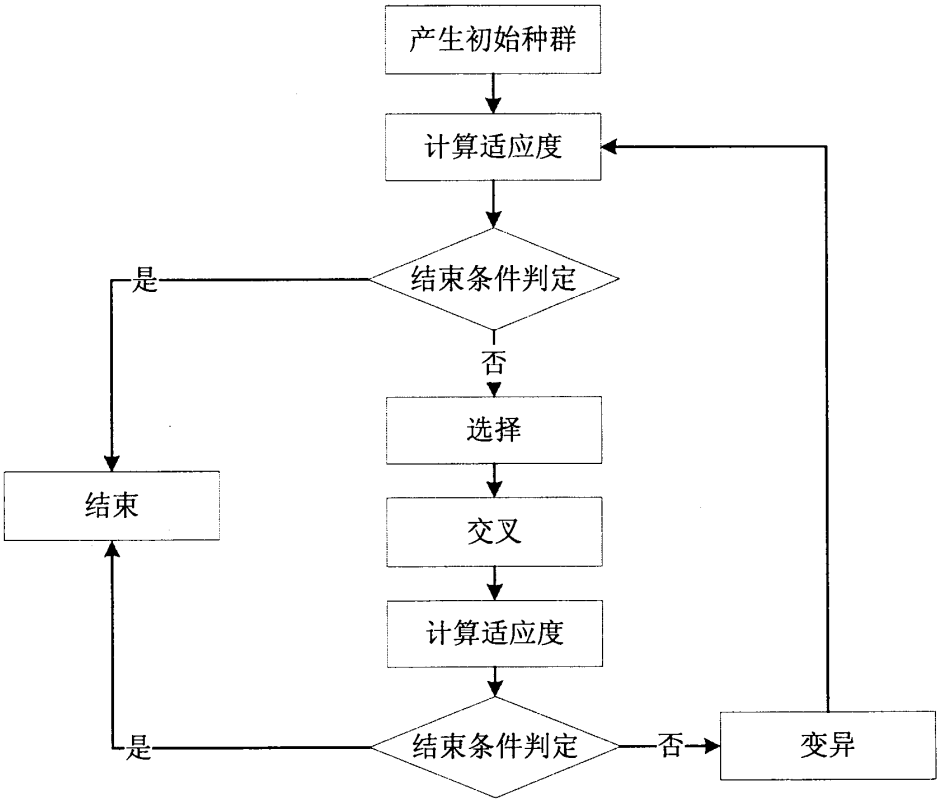


图 2-10 遗传算法流程图

遗传算法的下一步是通过交叉、变异两个遗传操作生成下一代个体，组成新的种群，新产生的后代个体继承了父个体的大部分特征。产生的后代个体再被选择作为父个体产生新的后代个体。这个过程使得新种群不同于初始种群，同时个体的适应度也会向较优的方向发展，初始种群中适应度较好的个体才被保存下来，其他个体都将被替代。虽然交叉和变异是遗传操作常用的方法，我们也可以在遗传算法中使用重组、迁移等方法。对变异概率、交叉概率和种群大小的研究也是很重要的，很小的变异概率也可能会对遗传结果产生较大影响。重组概率如果较高的话，可能导致遗传算法的早熟。而变异概率过高会错失一些较好的解决方案。但是还没有关于这些参数上限和下限设置的成熟理论。

遗传算法的终止条件有多种，设定进化次数；找到满足最低要求的解；人工检验等。上述方法可以单独使用，也可以多种方法结合使用。

遗传算法虽然有较广泛的应用，但是其自身存在着局限性。首先由于个体的适应度在每次迭代产生后代时都需要计算，当种群数量较大时，适应度的计算将耗费较多的时间。其次问题的复杂性不具有很好的可扩展性，因为变异操作个体的增多，伴随着搜索空间的指数增长。再者，与其他传递启发式方法相比，遗传算法的效率并不是最优的，甚至比其他算法的执行效率低很多。另外，遗传算法个体的选择太依赖适应度函数，然后没有严格的理论证明针对特定问题，选择何种适应度函数将是最优的。最为重要的是，遗传算法在解决动态变化的数量方面没有优势，因为在初始生成种群时，种群的大小已经确定。

2.4VoIP 流量识别方法

VoIP 流量是网络流量的一种，但是由于 VoIP 软件和流量特征与常规软件有所不同，决定对其流量的检测识别不用使用传统的识别方法，诸如基于端口的识别、基于特征码的识别以及基于流量数学特征的识别方法等。下面简单简介几种 VoIP 流量识别的常用方法。

2.4.1 基于主机行为特征的流量识别方法

基于主机行为分析的方法研究主机之间的关系，主机之间的通信首先是主叫与被叫之间建立连接关系[28]。通信双方建立连接时的数据交换信息保存在数据包头部中，提取数据包头部信息而不需要对负载数据进行分析可以识别流量类型。另外，识别某些类型软件的流量，可以通过综合分析服务器主机数量、客户数量以及一组通过应用程序相互连接的主机使用过的端口号实现。该方法同样适用于使用 P2P 协议的应用程序，这些应用程序实现一对主机之间的通信，而且同时使用 TCP 和 UDP 协议，在不同的主机上使用不同的端口号。

P2P 通信是不同于客户/服务器通信模式的，这种方法可以有效提取 P2P 协议软件特征[29][30]。但是如果一台主机上同时运行多个基于 P2P 协议的应用程序，基于主机行为识别方法的精确度不能得到保证。另外提取流量特征值并完成相关的计算，时间复杂度较高，对终端性能的要求也相对较高。

2.4.2 基于 SIP 和 H.323 协议的流量识别方法

基于协议的流量识别分析协议信令流，提取信令流的特征值，达到流量识别的目的[31]。SIP 协议和 H.323 协议是 VoIP 应用程序使用率较高的两个协议。对这两种协议的软件流量进行分析时，要结合端口号和流量特征，才能判断出具体的协议类型，进而识别到这两种协议软件的流量。

在前面有关 SIP 的介绍中，我们知道 SIP 请求和应答消息的格式特点，因此对

于 SIP 流量的识别，只要检测请求与应答消息。通过对流量数据包的分析，建立相应的格式规则，请求报文的规则可定义为如下正则表达式

$$(invite|register|cancel)sip[\x09-\x0d-~]*sip/[0-2]\[0-9]$$

相应的应答报文的正则表达式如下

$$sip/[0-2]\[0-9][1-5][0-9][0-9][a-zA-Z]*$$

对于满足上述正则表达式的请求和应答消息，可以认为是 VoIP 流量。

H.323 是 ITU-T 定义的标准协议栈[32]，它的呼叫信令共有五个阶段：接纳控制与呼叫建立、通信能力交换与模式设定、建立媒体通信逻辑信道、通信进行过程中的呼叫服务、呼叫结束。由于在通信建立过程中的数据传输涉及 TCP 和 UDP，通过识别流量的类型，结合端口识别方法可以判断流量是否为 VoIP 流量。

2.4.3 基于流量特征提取的识别方法

传统的流量识别方法中，有一类重要的载荷分析法。分析数据包头部信息，或者分析流量载荷[33][34]。这种方法的弊端在于，不能用于分析经过加密的流量。VoIP 流量正是经过加密的，而且是使用 AES 加密算法，破解后再分析是完全不可行的。因为数据包在网络上传递，涉及数据包平均长度、包到达时间间隔等特征，而且这些特征对于不同的软件表现不同，因此可以通过对这些特征的分析，达到识别流量的目的。在特征分析方面，朴素贝叶斯分类方法是比较常用的。

朴素贝叶斯方法是分析流量特征，无须检验数据包的具体内容，不仅能够防止数据内容的泄露，同时能够胜任加密数据的处理[35]。另外现在软件种类不断出现，新的软件的应用给流量识别工作也带来挑战，传统方法不能很好地对新增软件流量进行识别。朴素贝叶斯方法在这种情况下完全适用，只需通过对大量数据流的分析及特征提取，完全可以识别新增软件产生的流量。朴素贝叶斯方法在流量识别方面有很多的优点，但是其缺点也是明显的。首先，随时网络应用的增长，单纯依据数据包长度和包到达时间间隔来识别流量，其准确度得不到保障，容易造成误报和漏报。其次，朴素贝叶斯方法需要统计分析大量数据，执行周期较长，对于实时监控时的流量识别，不能及时给出结果。同时当数据量较小时，该方法也不能给出准确的分类结果。

2.5 本章小结

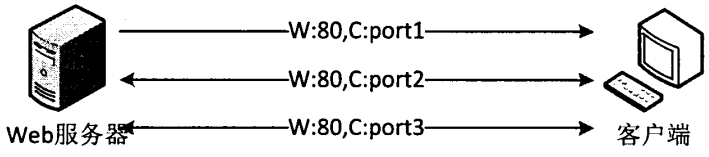
本章主要对与 VoIP 相关的网络协议 SIP、H.323、RTP/RTCP 和 Skype 等进行分

析。介绍机器学习的两种方法，特别是人工神经网络这种应用领域较多的算法。最后重点介绍了当前常用的几种 VoIP 流量识别方法，包括基于主机行为的流量识别方法、基于 SIP 和 H.323 协议、基于流量特征提取的识别方法，并且分析了每种方法的不足之处。

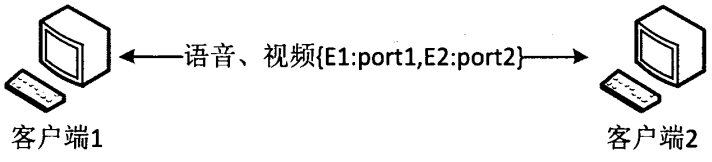
第三章 基于机器学习的 VoIP 流量识别

3.1 主机行为特征

主机可以运行不同类型的软件，由于软件结构和模型的不同，通信时表现的特征也不相同。对于网络软件来说，端口号是重要的资源，因为客户端必须通过端口号才能进行彼此通信。传统网络应用程序大部分是客户端/服务器模式。这种模式下，客户端与服务器的通信在多个端口下并行进行，这样可以提高通信效率。例如 Web 应用程序，IP 地址为 w 的服务器监听服务商品 80，IP 地址为 c 的客户端使用多个端口号与服务器的 80 端口通信，从服务器获取数据。这种方式被称为管道机制，如图 3-1(a)所示。在连接通信期间，服务器使用 80 端口，客户端使用的端口数量比服务器多。对于 VoIP 类型的 P2P 通信软件，用户通过认证后建立端到端的通信，传输语音和视频信息[36]，如图 3-1(b)所示。这种通信不属于管道通信，因为通信的主机使用的端口是一致的。



(a)Web 应用程序模型



(b)VoIP 应用程序模型

图 3-1 通信模型

分析主机行为特征，可以用五元组表示 $\{srcIP, srcPort, dstIP, dstPort, transport\ protocol\}$。其中 *srcIP* 和 *srcPort* 分别表示源端主机的 IP 地址的端口号，对应的目的端主机的 IP 地址和端口号为 *dstIP* 和 *dstPort*，通信协议使用元 $\{transport\ protocol\}$ 表示[17]。主机行为特征模型可以定义为

$|srcPort.num - dstPort.num| \leq t$ ，其中t是预定义的系统参数。当t的取值很小时，表示两个主机应用程序通信时使用端口号数量基本相同，符合 VoIP 软件通信特点。

3.2 流量行为特征

数据包长度和分组到达时间间隔是描述网络流量特征的两个主要参数，如表 1 所示。对于不同应用程序产生的流量，这两个值是不同的。为确保应用程序在不同网络环境下正常运行，并且保证较优的资源利用率，因此针对不同的软件这两个值的设置会有很大差异。

表 1 网络流量特征参数

特征参数		单位
数据包长度	分布情况	[最小(Bytes),最大(Bytes)]
	平均值	Bytes
	中值	Bytes
	变异	个
分组到达时间间隔	分布情况	[最小(ms),最大(ms)]

数据包长度是由源主机编码器的类型决定的，主要有两个类型，可变比特率和固定比特率。可变比特率虽然采样频率是固定的，样本大小会受到其他因素的影响产生较大变动，以获得较优的数据压缩质量。而固定比特率的样本大小是固定不变的，亦即流量数据包大小是固定的。相较而言，可变比特率的编码方法更具灵活性，但也更加复杂。VoIP 流量数据包包含可变比特率和固定比特率编码器编码的数据。

网络操作中，由排队、协商导致的网络碰撞，使得分组到达时间间隔趋向于平均值为T的随机值。VoIP 流量的分组到达时间间隔通常呈现波形连续分布，如图 3-2 所示为 QQ 语音分组到达时间间隔。连续的数据之间大小关系是交替出现的。为了量化表示，定义EL为分组到达时间间隔大值的评估值，ES为分组到达时间间隔小值的评估值，EL和ES是评估到达时间间隔的上限和下限计算方法如公式(3.1)和(3.2)所示，其中α权重系数，t_i表示第i时间点[17]。对邻接的时间进行分析时，需要满足每个公式后面的条件。

$$\begin{cases} EL_i = \alpha \cdot EL_i + (1-\alpha)t_i & t_i - ES_i > EL_{i-1} - t_i \\ EL_i = EL_{i-1} & others \end{cases} \tag{3.1}$$

$$\begin{cases} ES_i = \alpha \cdot ES_{i-1} + (1-\alpha)t_i & t_i - ES_i < EL_{i-1} - t_i \\ ES_i = ES_{i-1} & others \end{cases} \tag{3.2}$$

为了衡量 EL 和 ES 之间的差异，公式(3.3)定义两者之间的比例关系 R。在 VoIP 应用中，应确保 R 的值不能太大，因为 VoIP 应用对网络延迟很敏感，而且不能出现太多的碰撞。

$$R_i = \frac{ES_i}{EL_i} \tag{3.3}$$

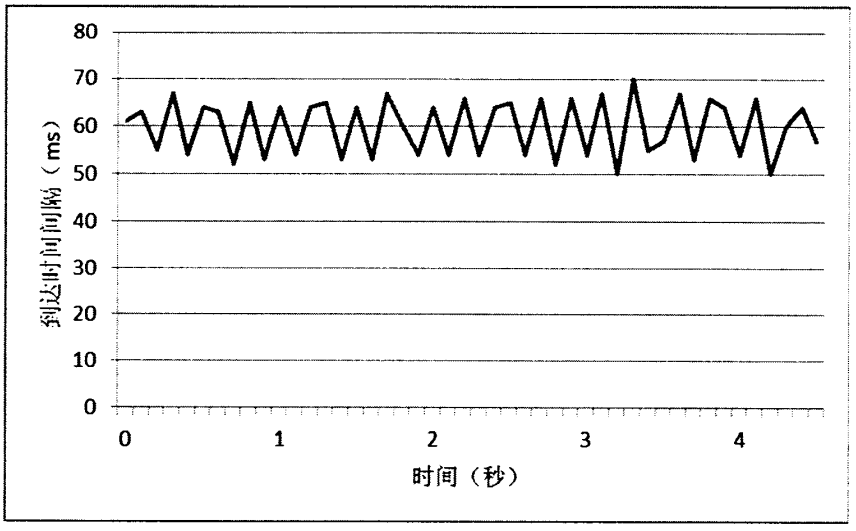


图 3-2QQ 语音分组到达时间间隔

3.3 机器学习流量识别

基于机器学习的流量识别方法主要有两个部分组成，即建立机器学习模型和使用机器学习进行分类。该方法使用到的数据有两种来源，一种是通过流量捕获程序采集主机产生的通信流量，另外一种途径是从 TSAT(TCP Statistic And Analysis Tool)官方网站 http://tstat.polito.it/traces-skype.shtml 下载 Skype 流量跟踪数据包[37]。由 3.1 的分析，根据流量包头文件中源端口和目的端口的差值，可以对流量数据进行预分类，提取特征值形成特征库文件。然后再选择适量的样本数据，进行特征提取，得到最优特征子集。再使用机器学习算法训练最优特征子集，输出机器学习分类结果。流程如图 3-3 所示。

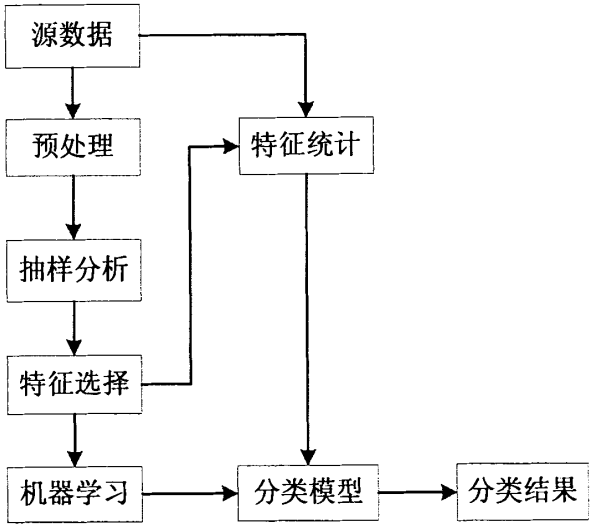


图 3-3 机器学习分类方法

3.3.1 流量特征选择

原始的流量数据是不能直接用于机器学习的，流量具有一些典型的特征，可以用参数化的方式表示这些特征。原始流量的特征总共有 246 个[12]，但是由于机器学习的输入参数的限制，如果将 246 个特征都作为输入，虽然可以提高流量识别的准确度，但是对于计算机是不可能实现计算的。因此需要选择其中关键的特征作为流量识别的依据。而且流量识别的准确度和可靠性在很大程度上是由特征选择决定的。

在特征选择方面，为了准确识别 VoIP 服务，应该选择较多的特征，但是对于流量识别的计算则不应如此。无效的特征不仅影响分类结果，同时增加计算复杂度。因此选择核心特征才能提高计算的准确性。

假设集合 A 是 VoIP 样本数据，集合 B 是非 VoIP 样本数据，核心特征的选择是依据两个样本之间的差别。设 x_i 表示集合 A 的一个样本，y_j 表示集合 B 的一个样本，$d(x_i, y_j)$ 表示样本间或样本内的差别，则集合 A、B 的平均差如公式(3.4)所示

$$L(x) = \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n} d(x_i, y_j) \tag{3.4}$$

其中 m 表示 VoIP 样本量，n 表示非 VoIP 样本量。如果 x_i 和 y_j 是同一集合中的样本，

那么 $d(x_i, y_j)$ 表示集合内部差。

对于特征 X_k，如果集合 A 和 B 的平均差值明显高于集合 A、B 各自的内部差值，那么特征 X_k 可以认为是核心特征。如果平均差值和 A、B 内部差值接近，也就是说差值不是很明显，那么可以认为 X_k 不是核心特征。

使用上述方法，对网络流量特征分析后，我们得到如表 2 所示的 20 个典型特征值。

表 2 20 个初始特征

特征
流量参数（时间间隔，分组数量，字节总数）
报文到达时间（平均值，最小值，最大值）
TCP/IP 控制字段大小（平均值，最小值，最大值）
总计报文数量（发出报文数量，接收报文数量）
负载大小（平均值，最小值，最大值）
TCP 域（应答字段）
传输层特征（协议，端口，{地址,端口}）
比例特征（接收/发送连接数量比例，上传/下载流量比例）

3.3.2 贝叶斯网络

贝叶斯分类算法是一种基于随机事件概率分布的统计分类方法，常用的朴素贝叶斯是一种有监督的学习方法，假定所有特征都是重要且相互独立的，依据成员特征概率总和，将每个对象与特定类别相关联[38]。该方法中的成员概率基于高斯曲线计算。但是对模型的假设以及特征重要性、相关独立性的假设并不是在任何情况下都有效。

已知一组流量 $x=(x_1, x_2, \cdots, x_n)$，其中流量 x_i 是 m 个元素的组合 $\{d_1^{(i)}, \cdots, d_m^{(i)}\}$，每个元素可以是数据也可以是离散值。对于网络流量 c，$d_j^{(i)}$ 是流量 x_i 的特征。例如可以表示流量 x_i 数据包平均到达时间间隔。监管贝叶斯分类方法，在训练数据的基础上建立统计模型，使得每个待处理的流量 y 都可以依据公式（3.5）的规则，被划分到特定的类别。

$$p(c_j)=\frac{p(c_j)f(y|c_j)}{\sum_{c_j}p(c_j)f(y|c_j)} \tag{3.5}$$

其中 $p(c_j)$ 表示相互独立的观察数据被划分为类别 c_j 的概率，$f(y|c_j)$ 是分布函数，分母表示正常化参数。

朴素贝叶斯方法的好处在于不需要复杂的搜索过程，只要独立对计算每个训练实体各个特征出现的频次，就可以估计出每个特征的概率估计值。在上面提到朴素贝叶斯方法理论基础的两个假设，这两个假设在真实网络环境中都是不能满足的，因此需要改进朴素贝叶斯方法，使之更适用复杂网络环境。

贝叶斯网络是对朴素贝叶斯方法的改进，使用概率图模型表示一组随机变量和变量之间的条件依赖关系。形式上贝叶斯网络由一个有向无环图和一个条件概率集合组成。图中的结点表示随机变量，可以是能够直接观测到的变量值，也可以是集合类型的变量，边表示随机变量间的条件依赖关系。

贝叶斯网络有一条非常重要的性质是不能忽略的。即网络中每个节点，它的直接前驱节点值被确定后，由前驱节点可以计算得到当前节点的值，与其他非前驱节点是相互独立的。一般情况下，对于集合 $x=(x_1,x_2,\cdots,x_n)$ 中的变量不是相互独立的，此时联合条件概率的计算方法如公式(3.6)所示

$$P(x_1,x_2,\cdots,x_n)=P(x_1)P(x_2|x_1)P(x_3|x_1,x_2)\cdots P(x_n|x_1,x_2,\cdots,x_{n-1}) \tag{3.6}$$

贝叶斯网络中每个节点的概率计算只与其直接前驱节点有关。对于任意随机变量，计算其联合概率分布的方法如公式(3.7)所示。其中 $Parents(x_i)$ 表示 x_i 的直接前驱节点的联合概率。对于没有前驱节点的节点概率依据经验计算。

$$P(x_1,x_2,\cdots,x_n)=\prod_{i=1}^{n}P(x_i|Parents(x_i)) \tag{3.7}$$

3.3.3 贝叶斯网络训练

贝叶斯网络训练的主要目的是为特定问题找到最优分类函数。在监督学习中，使用一组训练集训练网络学习能力，使函数适应给定的数据集合。通常的方法是定义损失函数 c，作用于输出值和目标值 $\{t_i\}$，或者作用于网络对某些训练输入评估得到的预测值 $\{p_i\}$ 与目标值 $\{t_i\}$，目标值是输入训练集的实际输出。预测值依赖所选

择的权重，因此可以计为有 $p_i = p_i(w)\,\forall i$。总体损失值可以通过公式(3.8)计算得到

$$L(w) = \sum_i c(p_i(w), t_i) \tag{3.8}$$

为了防止过拟合，需要降低权重损失值 L，同时服从约束条件 $\Omega(w)$。根据拉格朗日乘数原理，得到公式(3.9)，将降低损失值转化为降低公式(3.9)的值。

$$L(w) + \alpha\Omega(w) \tag{3.9}$$

完全贝叶斯方法根据贝叶斯理论给每个权重设置一个前驱，并在这些基于贝叶斯理论的未知因素之间构造后继节点。在这一点上，贝叶斯和完全贝叶斯是相似的。这里将对数似然函数作为损失函数，对数前驱类似于正则化矩阵。不管怎样，贝叶斯方法有其确定的优点，例如从后继节点权重误差线的生成。正则化矩阵的其他参数始终被选择，并且被计算作为模型选择的依据。

部分特征预处理需要前驱节点的输入作为训练数据。我们对线性特征曲线进行缩放，使得训练数据的平均值为 0，同时标准误差为 0.5。这么做是可行的，因为神经网络包含一个一阶非线性激活函数。这种缩放使得输入的变化服从类似的非线性变化，而且标准误差值为 0.5 已经被证明是最佳的。在网络训练中，这种转换会被存储起来，以便可以重新调整用于预测的输入测试集。

本文，我们在给定参数 $\phi = w(3)$ 的情况下，定义用于分类的似然函数，并引入最大后验概率(MAP)评估神经网络权重，这些权重最终被用于分类预测。权重在似然函数中是未知参数，因此为这些未知参数引入先验分布 $P(w)$。高斯先验分布是在神经网络中应用最多的先验分布模型。但是我们并不会特别重视某种特定网络而忽略其他网络，我们希望尽可能扩大权重选择的范围。高斯分布的形式表示为公式(3.10)

$$\propto \exp(\alpha F(w)) \quad F(w) = -\frac{w^2}{2} \tag{3.10}$$

前驱最大熵可表示为 $F(w) = |w|\log(|w|)$，极限值收敛于零比高斯方法更慢。相比高斯表达式，前驱最大熵是更合适的选择，因为它对较大的权重值的约束更少，而较大的权重值的训练需要建立精确的分类器。通过证明验证表明，前驱最大熵能够为很多问题提供较好的解决方案。

因此本文使用前驱最大熵方法，定义如公式(3.11)

$$P(w|\alpha)=\frac{1}{Z_s(\alpha)}\exp\alpha S(w) \tag{3.11}$$

其中正负权重的熵S的定义如公式(3.12)

$$S(w)=\sum_{w_{ij}\in w}(\psi_{ij}-2m-w_{ij}\log(\frac{\psi_{ij}+w_{ij}}{2m})) \tag{3.12}$$

其中$\psi_{ij}=(w_{ij}^2+4m^2)^{\frac{1}{2}}$。

上述定义包含两个参数m和α，m是默认级别并取值0.15。因为它起到的作用很小，因此我们不关注与这个参数有关的前驱依赖关系。参数α对训练过程影响很大，因此需要将它保留在计算公式中。贝叶斯定理可以用公式(3.13)表示

$$P(\phi|H,D)=\frac{P(\phi|H)P(D|H,\phi)}{P(D|H)} \tag{3.13}$$

由此推导出公式(3.14)

$$P(w|D,\alpha,H)\propto Z_S(\alpha)^{-1}\exp\alpha S(w)P(D|w,H) \tag{3.14}$$

那么神经网络分类似然函数计算方法如分式(3.15)所示

$$P(Class=j|\phi,I,H)=u_j=f(j,w,I) \tag{3.15}$$

公式(3.14)为似然函数训练集中每个流量$f(j,w,I)$的输出，如公式(3.16)

$$P(D|w,H)=\prod_i f(c_i,w,I_i) \tag{3.16}$$

由前驱分布的相关推理可以得到后继分布的计算公式(3.17)

$$P(w|D,\alpha,H)=\frac{Z_S(\alpha)^{-1}\exp\alpha S(w)[\prod_i f(c_i,w,I_i)]}{P(D|\alpha,H)} \tag{3.17}$$

公式(3.17)为网络中权重a值的后继分布。然后使用共轭梯度算法在网络环境中计算得到权重的最大后验概率值。算法初始工作是分析搜索权重和a空间，该工作的基础是a的值保持恒定，而且在w空间内评估最大后验概率。从a派生的值被评估后，a将减小，算法再次被重复执行。当后验概率足够大时，算法过程结束。

由此，给定一些训练数据，我们就可以推算出权重的可能取值。这种方法可以准确对互联网中采集的流量数据进行分类。

3.4 本章小结

本章首先介绍传统的流量识别方法。分别是基于主机行为特征的流量识别方法和基于流量行为特征的流量识别方法，这两种方法虽然能够对网络流量进行分类，但是存在较大的局限性。因此引入机器学习方法，利用其自学习能力，通过贝叶斯网络建立分类模型。

第四章 实验方法及结果分析

4.1 实验环境

使用 SIP Express Router（SER）和 SIPp 搭建基于 SIP 的 VoIP 环境。SER 是开源软件，可以作为客户端和服务器代理。SIPp 是用于生成 SIP 流量的专业测试工具。SIPp 和 SER 配合使用，可以仿真网络上的 SIP 活动。然后再使用 Wireshark 采集 VoIP 环境运行产生的流量，记录形成文件。这些文件用于贝叶斯网络的训练和流量识别分类原始数据样本。在流量生成过程中，使用 INVITE Flooder 软件生成 DoS 攻击流量。垃圾流量的存在可以很好地验证本文算法的优劣。贝叶斯网络的模拟，使用 Matlab 提供的贝叶斯类库实现。

4.2 实验数据集

本文使用的测试数据集主要有两方面的来源。第一是 4.1 节模拟环境采集的仿真文件。第二是从 TSAT(TCP Statistic And Analysis Tool) 官方网站 http://tstat.polito.it/traces-skype.shtml 下载的标准流量跟踪数据包。图 4-1(a)所示是实验环境 Skype 流量跟踪数据包，4-1(b)是真实网络环境 Skype 流量跟踪数据包。

Skype E2E Traces									
	Calls type	L4-protocol	Voice codec	Info codec	Start time	End time	L4-payload size per packet	Download Trace	File size
Skype	E2E voice only	UDP	GIPS iSAC	For more information about the iSAC codec, look here	2006-06-26 10:37:46 CET	2006-06-26 10:56:22 CET	All	Download	8.3 MB
Skype	E2E voice only	UDP	SVOPC	For more information about the SVOPC codec, look here	2008-01-11 12:37:00 CET	2008-01-11 12:42:15 CET	All	Download	4.6 MB
Skype	E2E voice only	UDP	GIPS iLBC	For more information about the iLBC codec, look here	2006-06-14 10:25:08 CET	2006-06-14 10:41:52 CET	All	Download	6.0 MB
Skype	E2E voice only	UDP	GIPS iPCM-wB	For more information about the iPCM-wb codec, look here	2006-06-14 11:38:16 CET	2006-06-14 11:52:55 CET	All	Download	24 MB
Skype	E2E voice only	UDP	ITU-T G729	For more information about the G729 codec, look here	2006-06-14 11:21:56 CET	2006-06-14 11:37:01 CET	All	Download	3.8 MB
Skype	E2E voice only	UDP	EG711 (A-law)	EG711 is an Enhanced version of the ITU-T G.711 codec	2006-06-15 10:51:26 CET	2006-06-15 11:06:44 CET	All	Download	22 MB
Skype	E2E voice only	UDP	EG711 (u-law)	EG711 is an Enhanced version of the ITU-T G.711 codec	2006-06-15 10:35:26 CET	2006-06-15 10:49:41 CET	All	Download	22 MB
Skype	E2E voice only	UDP	PCM (A-law)		2006-06-15 11:22:35 CET	2006-06-15 11:37:18 CET	All	Download	16 MB
Skype	E2E voice only	UDP	PCM (u-law)		2006-06-15 11:07:53 CET	2006-06-15 11:21:28 CET	All	Download	16 MB
Skype	E2E voice only	TCP	GIPS iSAC	For more information about the iSAC codec, look here	2006-09-25 14:21:38 CET	2006-09-25 14:37:01 CET	All	Download	7.3 MB
SkypeOut Traces									
	Calls type	L4-protocol	Voice codec	Info codec	Start time	End time	L4-payload size per packet	Download Trace	File size
Skype	SkypeOut call	UDP	G729	In this case, a PSTN-user located in Torino, was contacted	2006-09-26 09:09:56 CET	2006-09-26 09:25:06 CET	All	Download	3.8 MB
Skype	SkypeOut call	TCP	G729	In this case, a PSTN-user located in Torino, was contacted	2006-05-26 12:40:54 CET	2006-05-26 12:56:14 CET	All	Download	4.8 MB

(a) 实验环境 Skype 流量跟踪数据包

Skype Traces					
	Type	L4-protocol	L4-payload size per packet	Download Trace	File size
Skype	E2E voice only and voice+video calls	UDP	max 38 Byte	Download	1.3 GB
Skype	Skype Out calls	UDP	max 38 Byte	Download	67 MB
Skype	Signaling connections only	UDP	max 38 Byte	Download	2.2 GB
Skype	E2E and SkypeOut calls	TCP	max 26 Byte	Download	55 MB

(b) 真实网络环境 Skype 流量跟踪数据包

图 4-1TSAT 官方 Skype 流量跟踪数据包

监测获得的数据是监控互联网一个站点不同时期产生的流量，这个站点必须方便用户访问，而且能够提供 1000 个用户通过全双工光纤连接到互联网。全双工连接通信中产生的每个流量数据，都会被监控系统记录保存。这些流量可能是一种应用程序产生，也可能是不同的应用程序在同一时间段内产生的，包括网页服务、游戏、邮件服务、流媒体传输以及下载软件等。另外，实验需要多组数据，因此在不同的时间段分别运行监控器，分别保存多个流量文件。这些数据采集保存后作为实验的原始数据，供特征提取有分类使用。

表 1 软件分类

分类	应用程序
数据库	Oracle，SQL Server
邮件	IMAP，SMTP，POP3
P2P	Skype
游戏	Warcraft
多媒体	PPStream，Realplayer

不管是通过监控器采集的数据还是 TSAT 官方数据，首先需要进行预处理。预处理可以按不同的分类方法将流量划分为不同的分组。首先，根据网络协议进行分类，可以将数据流量划分为 TCP 流量和 UDP 流量。3.1 节中介绍到主机行为特征时涉及到五元组 $\{srcIP, srcPort, dstIP, dstPort, transport\ protocol\}$，该信息保存在数据包的包头中，对包头分析提取 *transport protocol* 字段，得到应用程序协议，由此可以确定协议类型。其次，可以根据流量特征分类。提供数据包包头中有关分组数据大小和分组到达时间的相关信息，使用 3.2 节介绍的方法可以分离出 VoIP 流量和非 VoIP 流量。但是这种分类方法精确度很低，因为当新研发的软件表现出与当前软件不同的特征时，是不能被检测到的。另外，人们根据应用程序的功能，归纳出几种基本类型的应用程序，如表 1 所示。但是这种分类方法有很大的局限性，因为根据功能分类涉及不到所有的软件。而且出现新的软件时，需要更新列表。

4.3 分类方法评价指标

为了准确评价分类方法的性能，本文使用精度(accuracy)、查准率(precision)和查全率(recall)作为评价指标[39]。以上三个指标与 TP、FP 和 FN 相关，其中：

-True Positive(TP)，被模型预测为正的正样本

-False Positive(FP)，被模型预测为正的负样本

-False Negative(FN)，被模型预测为负的正样本

精度(公式 4.1)主要用于评估分类器的准确性，是指对于给定的测试数据集，分类器正确分类样本数与总数之比。

$$accuracy = \frac{\sum_{i=1}^{n} TP_i}{\sum_{i=1}^{n} TP_i + \sum_{i=1}^{n} FP_i} \times 100\% \tag{4.1}$$

查准率公式(4.2)和查全率（公式 4.3）是对分类结果的整体效果进行评价。

$$precision = \frac{TP}{TP + FP} \times 100\% \tag{4.2}$$

$$recall = \frac{TP}{TP + FN} \times 100\% \tag{4.3}$$

4.4 实验结果分析

在实验原始数据中选取 19946 个流数据包，其中属于 P2P 类型的流数据包有 10166 个，其他非 P2P 流数据包计为 9780 个。基于贝叶斯网络的分类方法，对原始数据分类处理后，能够分离得到 P2P 数据和非 P2P 数据（记为 NP2P）。

4.4.1 分类检测结果

对 19946 个数据包使用贝叶斯网络分类方法后得到各类型流量统计结果如表 2 所示。由表可以看出，对 P2P 流量的检测精度达到 94.7。之所以比其他类型的流量检测精度稍低，主要原因是训练阶段样本选择较少。将训练样本增多后的结果如表 3 所示。

表 2 检测结果分类

类型	数量	所占比例	精度
P2P	9551	0.48	94.7
数据库	2248	0.11	97.6
邮件	1482	0.07	95.4
www	5852	0.29	93.8
其他	813	0.05	-
总计	19937		95.4

表 3 调整训练样本容量后的检测结果分类

类型	数量	所占比例	精度
P2P	10044	0.50	97.2
数据库	2248	0.11	97.6
邮件	1482	0.07	95.4
www	5852	0.29	93.8
其他	320	0.03	-
总计	19937		96.1

由表 2 和表 3 的对比可以看出样本大小的选择对结果有较大影响。图 4-2 所示是训练样本容量的选择和检测精度之间的关系。由图可以看出精度是随着样本容量的增加而逐渐增长的。这是因为当样本容量较小时，特征不能被准确提取出来，在分类时容易造成分类错误或漏分的情况，分类精度较低，且多次实验时精度变化范围较大。表 2 和表 3 中被划分为其他类的流量就属于漏分的情况。当样本容量达到一定大小时，每种类型应用程序的特征都可以被准确提取，为分类提供可靠的依据，因此精度会较高，且波动范围较小。图 4-3 所示为样本容量与精度误差之间的关系。由图可以看出，当样本容量较小时，多次实验得出的精度的波动范围比较大。明显的当样本容量为 100 时，多次实验的精度波动范围在是 3.2。随着样本容量的增长波动范围随之减小，当样本容量达到 600 时，波动范围是 0.8 左右。但是当样本容量达到一定值时，波动范围的变化会处于稳定状态。因此考虑检测精度、稳定性和计算量的情况下，应适量选择样本容量。对于稳定性要求不高但要求实时检测的情况，可以适量降低样本容量，降低计算量，满足实时性。

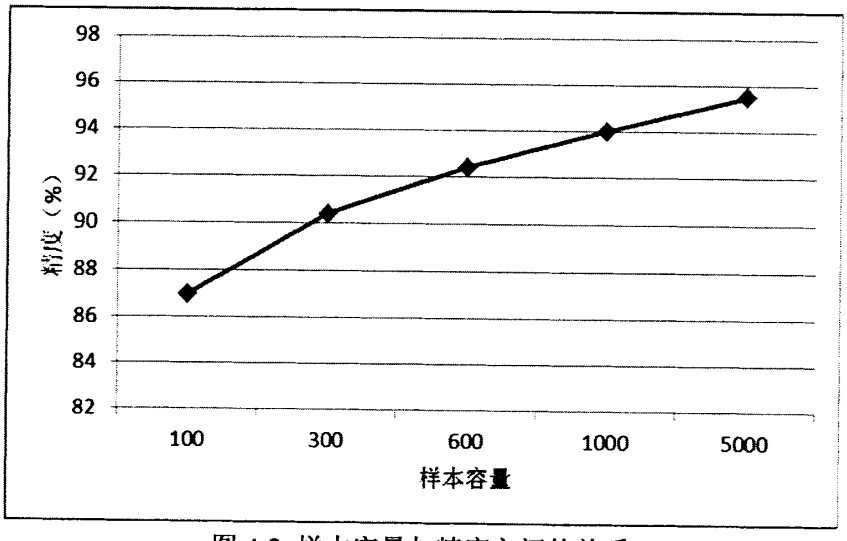


图 4-2 样本容量与精度之间的关系

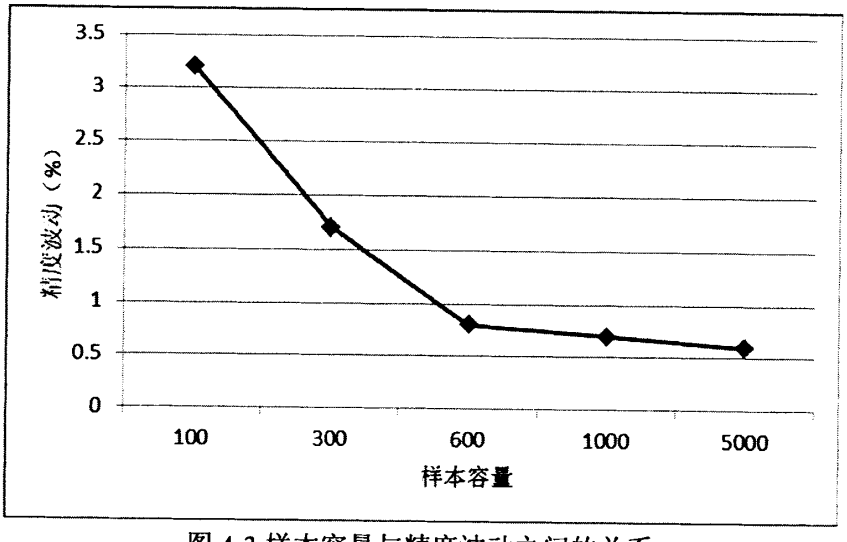


图 4-3 样本容量与精度波动之间的关系

4.4.2 不同机器学习算法的比较

本节比较在相同条件下，对同一样本数据进行分类的不同贝叶斯方法的实验结果，如表 4 所示。

表 4 各种分类算法的对比

分类算法	精度	查准率		查全率		分类时间
		P2P	NP2P	P2P	NP2P	
朴素贝叶斯	0.685	0.694	0.292	0.978	0.021	2
反向传播神经网络	0.699	0.697	0.986	0.996	0.015	490
贝叶斯网络（K2）	0.951	0.959	0.930	0.970	0.907	7
贝叶斯网络（遗传算法）	0.957	0.960	0.952	0.978	0.900	1230
贝叶斯网络（模拟退火）	0.958	0.960	0.950	0.980	0.902	3400
贝叶斯网络（爬山）	0.951	0.959	0.930	0.970	0.905	14
树增强型朴素贝叶斯网络	0.959	0.963	0.947	0.978	0.915	14
贝叶斯网络	0.961	0.965	0.953	0.980	0.919	17

从表 4 的对比结果可以发现，在查全率方面，朴素贝叶斯方法和反向传播神经网络方法的值都比较低，分别为 0.021 和 0.015，造成这种现象的原因在于这两种方法把非 P2P 流量误判为 P2P 流量，导致分类精度下降。在精度方面，贝叶斯网络是最高的，这是因为算法的改进去除非核心特征，引入核心特征。虽然神经网络训练方法的改进提高了检测精度，但是复杂度也随之提高，运行算法所需的时间也相应较长，因此会出现模拟退火算法的贝叶斯网络运行时间长达 3400 秒。综合考虑，贝叶斯网络(K2)、树增强型朴素贝叶斯网络和贝叶斯网络(BAN)在精度、查准率和查全率方面都较高，而且运行时间较短，是比较理想的分类算法。

4.4.3 特征参数数量与精度之间的关系

3.3.1 节中我们提到作为贝叶斯网络输入参数的特征个数共计 246，其中有些是比较核心的特征，可以用来作为输入参数，而有些特征作为输入参数不仅不能提高分类精度，反而会使分类精度降低。同时减少参数个数还可以降低训练时间，使得算法也适用实时网络环境检测。

多层感知神经网络中，初始输入层节点权重决定输入数量，而且影响下一层节点活性。如果某一特定节点的所有权重都为零，那么输入对下层节点的不产生影响。也就是说输出结果的分类概率不受该节点的影响，因此该节点在整个分类网络中基本不起任何作用。反之，如果给定节点的权重比其他输入节点的权重大很多，输出结果分类概率有可能只与该节点有关，此时只需考虑该节点，其他节点在整个网络中作用很小。因此我们定义公式(4.4)

$$Signal(i) = \sum_{w_{ij} \in input} | w_{ij} | \tag{4.4}$$

其中计算得到的和，是网络第一层与输入节点关联的所有权重相加。

通过对公式(4.4)的观察可以得到如下性质：小部分输入相对其他输入对结果的影响较大，这一性质类似于偏态分布。神经网络的输入数量不同时，可以利用这一性质作为标准，排除影响较小的输入。通过对 2441 个输入流的分析计算，我们得到如图 4-4 所示的结果。

从图中我们可以看出，输入数量小于 10 时，精度在 50%以下，不能满足检测要求。随时输入数量的增加，当达到 40 左右时，精度稳定在 92 左右。只有当输入数量达到 128 时，精度达到 95.3。而且继续增加输入，并不会显著提高精度。这说明当输入数量是 128 时，得到的精度值是比较理想的，而且计算 128 个输入所需的时间复杂度相对也较低。

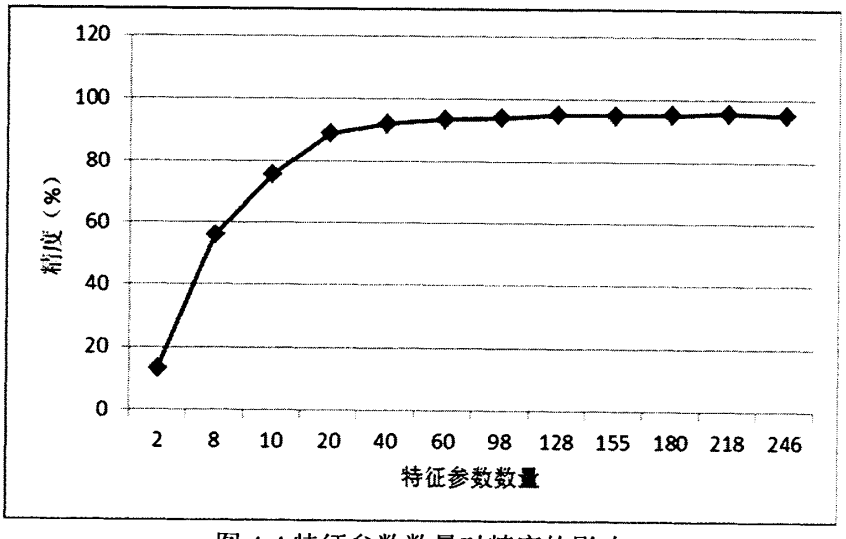


图 4-4 特征参数数量对精度的影响

4.5 实验总结

通过以上实验结果及分析，我们可以看到，本文提出的机器学习贝叶斯网络算法是有效的，精度可以达到 97%。而且相对其他流量识别方法，该方法精度、查全率和查准率都较高。在运行时间方面，该方法保持在 14 秒左右。同时保证较低运行时间的基础上，确保了较高的检测准确度。对特征参数的个数进行分析得到，当特征参数个数选择为 128 时，得到的计算结果精度较高，并且同时保证计算较低的计算时间。

4.6 本章小结

本章介绍实验所需数据的来源，对实验结果进行分析，以精度、查准率和查全率评估算法有效性，并与其他贝叶斯算法进行比较，说明贝叶斯网络分类的优势。最后通过对输入数量和分类精度之间关系的分析，确定 128 是贝叶斯网络最佳特征参数数量。

第五章 总结和展望

5.1 全文总结

VoIP 技术的发展，给网络流量检测技术带来挑战的同时也带来技术改进的机遇。VoIP 的广泛应用，给用户提供了快捷方便且价格便宜的语音、视频通信服务，特别是最近无线网络的发展，使得用户不受地址位置的限制，能够随时使用该服务。VoIP 给用户带来方便的同时，由于非法用户的存在，服务质量受到威胁，因此流量识别技术应运而生。

对 VoIP 流量识别的研究，从传统利用主机行为特征和流量行为特征进行流量分析识别，发展到将人工智能算法引入 VoIP 流量识别的应用中。本文主要利用贝叶斯网络方法实现 VoIP 流量识别，相较于传统的分析识别方法，该方法在保证分类精度的同时，运行算法的时间复杂度较低。因此该方法不仅可以运行于离线流量识别，也可以用于实时检测系统的流量识别。

5.2 技术展望

本文提出的机器学习贝叶斯网络方式实现 VoIP 流量识别，虽然有效且准确地解决了离线和在线的流量分类，但是作者相信通过优化特征选择以及选择更高效的算法，检测过程可以进一步改进。与此相关的优化方案实现和时间空间复杂度简化，是本题目进一步研究的方向。

参考文献

[1] Hilt V, Hari A, Hofmann M. An efficient and robust overlay routing scheme for VoIP[C]. Information, Communications and Signal Processing, 2005 Fifth International Conference on. IEEE, 2005: 508-512.

[2] VoIP-Info web site, http://www.voip-info.org/

[3] Tom-Skype web site, http://skype.tom.com

[4] Moore A, Zuev D, Crogan M. Discriminators for use in flow-based classification[M]. Queen Mary and Westfield College, Department of Computer Science, 2005.

[5] Dwivedi H. Hacking VoIP: protocols, attacks, and countermeasures[M]. No Starch Press, 2008.

[6] Forman G. An extensive empirical study of feature selection metrics for text classification[J]. The Journal of Machine Learning Research, 2003, 3: 1289-1305.

[7] Choi Y. On the accuracy of signature-based traffic identification technique in IP networks[C]. Broadband Convergence Networks, 2007. BcN'07. 2nd IEEE/IFIP International Workshop on. IEEE, 2007: 1-12.

[8] Kang H J, Kim M S, Hong J W K. A method on multimedia service traffic monitoring and analysis[M]. Self-Managing Distributed Systems. Springer Berlin Heidelberg, 2003: 93-105.

[9] Auld T, Moore A W, Gull S F. Bayesian neural networks for internet traffic classification[J]. Neural Networks, IEEE Transactions on, 2007, 18(1): 223-239.

[10] Du P, Abe S. Detecting DoS attacks using packet size distribution[C]. Bio-Inspired Models of Network, Information and Computing Systems, 2007. Bionetics 2007. 2nd. IEEE, 2007: 93-96.

[11] Ehlert S, Petgang S, Magedanz T, et al. Analysis and signature of Skype VoIP session traffic[J]. 4th IASTED International, 2006.

[12] Raahemi B, Zhong W, Liu J. Peer-to-peer traffic identification by mining IP layer data streams using concept-adapting very fast decision tree[C]. Tools with Artificial Intelligence, 2008. ICTAI'08. 20th IEEE International Conference on. IEEE, 2008, 1: 525-532.

[13] 杨国良. 国际 VoIP 流量特征分析[J]. 2007.

[14] 王振华, 王攀, 张顺颐. 基于综合统计特征的 Skype 流量分析与识别[J]. 南京邮电大学学报 (自然科学版), 2006, 1: 000.

[15] Jun L, Shunyi Z, Ye X, et al. Identifying Skype traffic by random forest[C]. Wireless

Communications, Networking and Mobile Computing, 2007. WiCom 2007. International Conference on. IEEE, 2007: 2841-2844.

[16] 王蕊, 张顺颐. 基于 P2P 的 Skype 与常规业务的流量分析和识别[J]. 通信技术, 2007, 40(05): 42-45.

[17] Li B, Ma M, Jin Z. A VoIP traffic identification scheme based on host and flow behavior analysis[J]. Journal of Network and Systems Management, 2011, 19(1): 111-129.

[18] 张登银, 通信技术, 孙精科, 等. VoIP 技术分析与系统设计[M]. 人民邮电出版社, 2003.

[19] ITU-T H.323. SERIES H: AUDIOVISUAL AND MULTIMEDIA SYSTEMS. 2009

[20] 陈蕾. H. 323 网守地址解析和路由管理的研究与实现[D]. 重庆大学, 2006.

[21] IETF RFC3261. SIP: Session initiation Protocol v.2.0. Rosenberg J, Schulzrinne H, Camarillo G, et al, 2002

[22] RFC 3550. RTP: A Transport Protocol for Real-Time. Schulzrinne H, Casner S, Frederick R, Jacobson V, 2003

[23] Erman J, Mahanti A, Arlitt M. Qrp05-4: Internet traffic identification using machine learning[C]. Global Telecommunications Conference, 2006. GLOBECOM'06. IEEE. IEEE, 2006: 1-6.

[24] Erman J, Mahanti A, Arlitt M, et al. Offline/realtime traffic classification using semi-supervised learning[J]. Performance Evaluation, 2007, 64(9): 1194-1213.

[25] Dedinski I, De Meer H, Han L, et al. Cross-layer peer-to-peer traffic identification and optimization based on active networking[M]. Active and Programmable Networks. Springer Berlin Heidelberg, 2009: 13-27.

[26] Erman J, Arlitt M, Mahanti A. Traffic classification using clustering algorithms[C]. Proceedings of the 2006 SIGCOMM workshop on Mining network data. ACM, 2006: 281-286.

[27] McGregor A, Hall M, Lorier P, et al. Flow clustering using machine learning techniques[M].Passive and Active Network Measurement. Springer Berlin Heidelberg, 2004: 205-214.

[28] Xu K, Zhang Z L, Bhattacharyya S. Profiling internet backbone traffic: behavior models and applications[C]. ACM SIGCOMM Computer Communication Review. ACM, 2005, 35(4): 169-180.

[29] Okabe T, Kitamura T, Shizuno T. Statistical traffic identification method based on flow-level behavior for fair VoIP service[C]. VoIP Management and Security, 2006. 1st IEEE Workshop on. IEEE, 2006: 35-40.

[30] Karagiannis T, Broido A, Faloutsos M. Transport layer identification of P2P traffic[C]. Proceedings of the 4th ACM SIGCOMM conference on Internet measurement. ACM, 2004: 121-134.

[31] Van Der Merwe J, Caceres R, Chu Y, et al. Mmdump: A tool for monitoring Internet multimedia traffic[J]. ACM SIGCOMM Computer Communication Review, 2000, 30(5): 48-59.

[32] 傅闪斌, 谭成翔, 汪海航. 基于 H. 323 的 VoIP 监测系统的研究与设计[J]. 计算机安全, 2007, 1: 8-10.

[33] 鄢汉科. 基于对称分段流量统计特征的 VoIP 流量识别[D]. 广州: 华南理工大学, 2010.

[34] 唐香兰. VOIP 的识别与流量分析[D]. 北京邮电大学, 2010.

[35] Moore A W, Zuev D. Internet traffic classification using bayesian analysis techniques[C]. ACM SIGMETRICS Performance Evaluation Review. ACM, 2005, 33(1): 50-60.

[36] Internet Assigned Numbers Authority, http://www.iana.org

[37] Skype trace web site, http://tstat.polito.it/traces-skype.shtml

[38] Erman J, Mahanti A, Arlitt M, et al. Semi-supervised network traffic classification[C]. ACM SIGMETRICS Performance Evaluation Review. ACM, 2007, 35(1): 369-370.

[39] Li W, Moore A W. A machine learning approach for efficient traffic classification[C]. Modeling, Analysis, and Simulation of Computer and Telecommunication Systems, 2007. MASCOTS'07. 15th International Symposium on. IEEE, 2007: 310-317.

在校期间发表的论文、科研成果等

发表论文

[1]肖诗松，冯辉，张峰，邹香玲. An Improved Algorithm Based on AC-BM Algorithm. 2012 计算机科学与信息、安全工程国际学术研讨会（CAISSE 2012）

[2]Xiao Shisong, Zou Xiangling, Zhang Feng, Feng Hui. Energy-based Cluster Partition Method in Wireless Sensor Networks, 2012 International Conference on Computational and Information Sciences(ICCIS2012), August 2012 Volume 3,912-915

参与项目

[1]2010.12-2011.5 参与湖北省 0-6 岁残疾儿童康复工程项目的设计与实现

[2]2011.7-2011.12 参与云计算机集群服务器管理监控系统的研发

致谢

三年的硕士生活从我们指尖，匆匆的滑过。美好的校园生活就要离我而去，迎来的是一个新的起点。时光的匆匆，远超了我的想象。三年收获了很多，不管是在学习、生活还是工作方面，都使自己上了一个崭新的台阶，马上就要离开美丽的华师，自己的母校。春天的华师，旧叶慢慢的飘落，新叶长出枝头，也许正标志着距离我们离开的步伐越来越近。春风扫着落叶，面对昔日入校时的样子，自己已经成长了很多。我即将离开美丽的校园生活，不免心生眷恋。

在自己的硕士论文即将落定之时，我要首先感谢我的导师肖诗松老师，肖老师严谨的科研作风，在培养学生的动手能力方面有很好的建树。肖老师三年来细心的培养，使我具备更加扎实的功底去面对即将到来的工作。另外还要感谢肖老师三年来对自己的关怀、理解和帮助。在肖老师的带领下，使自己有一个更好的学习和生活环境。

感谢计算机网络研究所肖德宝教授、崔建群教授以及赵尔敦教授的指导，使我涉猎的专业知识更加广泛。

感谢我的父母和我的弟弟妹妹，是他们给予了我学习的动力，不管是在精神还是在物质上都给予了我最亲，最爱的支持。

感谢我的室友董波、郝亨庚、严盟。是他们陪伴着我度过了三年的美好时光，是他们以宽容的态度，支持和理解我所做的一切。

感谢我的同门邹香玲、冯辉等。感谢大家对我的关怀和支持，很高兴在这个大家庭里学习生活。

感谢网络与通讯研究所的赖敏财、陈传河、齐勇强、相星星、沈青、董波。感谢他门一直陪伴着我走完自己的研究生生涯。

感谢好朋友王青、夏康、朱聚豹、朱松、李亚南等，感谢他们的支持与帮助。

感谢所有帮助过我，关心过我的人。谢谢你们的鼎力支持。

最后，由衷的感谢各位评审专家对我论文的批评与指正。

张峰

2013 年 5 月于桂子山