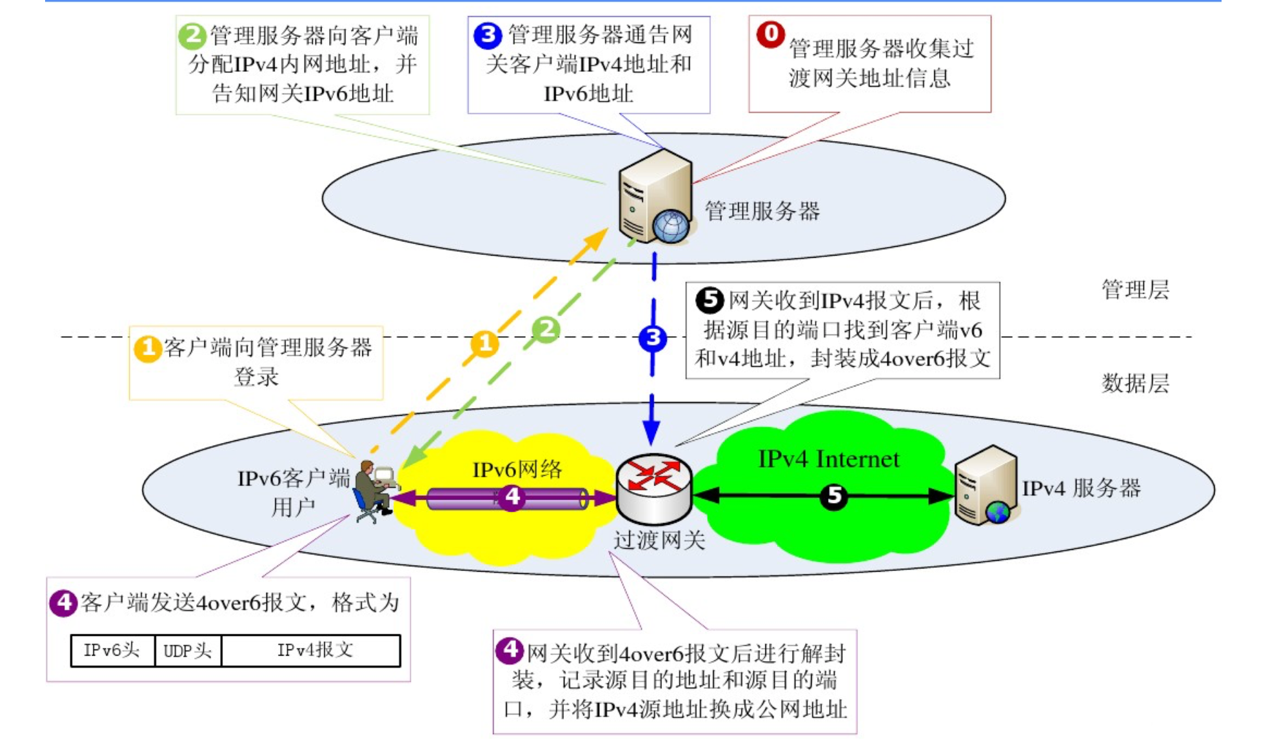
**基于机器学习的流量数据分析**

清华大学 计算机科学与技术系

队员：郭诚 李映辉 指导老师：徐明伟

1. **问题背景**

IPv6是下一代互联网协议，它的设计用以解决IPv4中的许多问题（最重要的一个就是解决了IP地址不够分配的问题）为目标。但是，在互联网大规模支持IPv6之前，IPv4到IPv6如何平滑过渡是需要人们仔细考虑的一个问题。例如，由于互联网服务提供商的现有网络设备大部分都只支持IPv4，我们有必要设计一种机制允许IPv6用户访问历史遗留的IPv4服务，而IPv4 over IPv6就是这样的一种机制。其大致原理是采用了“IPv4 over IPv6”隧道方案来实现，IPv6网络中的用户在希望访问IPv4服务时，将IPv4包封装在IPv6包中发送给一台能够直接访问IPv4服务的服务器。该服务器将进行实际IPv4资源的访问，然后再将得到的IPv4响应封装在IPv6包中发送回用户。为了能够通过IPv4地址将IPv4响应返回给相应的用户，IPv4 over IPv6服务器需要维护一个IPv4和IPv6地址的映射表并为需要IPv4 over IPv6服务的用户分配IPv4地址。原理大致如下图所示：



“马上6”平台是中国目前最全的IPv6资源网站，为广大用户提供最简单快捷、免费高速的IPv6网络访问IPv4网络的接口。因为其提供了快捷高速的IPv4 over IPv6接口，使得用户无需其他操作即可在IPv6网络中访问IPv4的资源。由于“马上6”的诸多好处，其产生的数据流量也非常大，在大数据时代，流量正在迅速增加，为了更好地优化其服务，我们有必要对其数据流量进行分类分析，利用新兴的机器学习算法来对“马上6”产生的数据流量进行分析，以达到优化其流量服务的目的。

1. **项目简介**

本项目与“马上6”公司方面取得了联系与合作，由清华大学计算机科学与技术系计算机网络研究所徐明伟老师担任指导老师，团队成员由清华大学计算机科学与技术系两位本科生——郭诚、李映辉组成。整个项目大致思路为，利用我团队自己开发的简易抓包程序在“马上6”的流量入口交换机上进行抓包获取特定时间段内的流量数据（IPv4、IPv6）并将其保存下来。对保存下来的数据利用已有的机器学习分类算法、聚类算法进行分析，从而从中分析出我们需要的“大小流信息”、“时间敏感性”、“存储敏感性”等要素，有了这些要素也进一步有利于“马上6”服务的优化与发展，同时也将网络行为检测分析领域与机器学习领域进行了一次交叉融合，利用机器学习方法的监督、半监督甚至是无监督学习等策略来提高网络行为检测的准确性和有效性。未来计划将项目落地到真实具体的“马上6”流量中，实现从“offline”到“online”的转变，真正实现实时的网络流量检测分析。

1. **项目方案**
2. **流量数据获取阶段**：

本阶段利用我团队自行开发的pcap抓包程序（即将抓取的网络流量数据包保存为普遍接受公认的pcap格式，方便后续IP地址匿名化处理）在“马上6”公司提供的交换机上进行抓包并存储到服务器上。初步抓包时间为一个小时的IPv4数据包，用以实现本项目提出的offline机器学习下的网络流量检测分析。运行环境为Centos，代码使用C编写，基于pcap库开发。

1. **包特征提取阶段**：

第一阶段获取的原始pcap数据文件包含了一部分分类所需的流量信息，如包长度、流量方向等，但是另外的一些信息需要进行特征提取，对原始数据进行预处理才可获得，如包间隔等，这些信息在流量分类中是很重要的。第二阶段的工作就是根据我们的预备方案在原始数据集上提取特征，整理出机器学习算法的输入数据集。

1. **IP地址匿名化阶段**：

IP流数据中有网络用户的IP地址、通信内容等隐私信息，如果不加处理或处理不当就将IP流数据向外界公布，势必会侵犯网络用户的隐私权或商业秘密。因此，我们的项目中涉及的“马上6”流量数据包也需要经过IP匿名化处理，即IP流净化处理。IP 流净化的指导原则是最大程度保持流数据 原来特性不改变和最大程度保护网络用户的隐私，主要进行以下三个方面的工作：首先，将IP地址匿名化，其次，消除数据中的用户通信内容，最后，重新计算IP头中的校验和。本项目中，使用基于crypto-PAn算法进行IP地址匿名化的C库——“CryptopANT”（written by Yuri Pradkin, University of Southern California, Information Sciences Institute, CA.）来进行我们的IP地址匿名化服务。其满足的功能包括：“IPv4 class awareness”（IPv4地址类感知）、“ Optional partial anonymization”（可选择部分匿名化）、“ IPv4 IPv6 MAC encrypt”（支持IPv4、IPv6、MAC地址加密）、“ decryption”（支持反匿名化）、“ Key generation”（支持自动生成密钥）。基于以上功能，我团队认为“CryptopANT”开源工具是满足我们的需求的。

1. **机器学习算法实现阶段**：

该阶段，我们预备使用多种机器学习算法对流量数据进行分类，根据前期论文调研的结果，准备采用以下几种算法：

**1、AutoClass：**

AutoClass是一种无监督的聚类算法，是基于概率模型的算法，可以人工指定类别数并自动实现数据的聚合。它是较早应用于流量分类的ML算法之一，在一些文章里被作为baseline实现。

**2、K-Means：**

K-Means算法是最快速和有效的聚类算法之一，是一种基于分区的算法，将其用在流量分析中也能起到不错的效果。

**3、DBSCAN：**

DBSCAN算法，全称是Density-Based Spatial Clustering of Applications with Noise，是一个比较有代表性的基于密度的聚类算法，能够把具有足够高密度的区域划分为簇，并可在含有噪声的空间高维数据集中发现任意形状的聚类。实验证明，DBSCAN算法虽然准确率略低于AutoClass算法，但是在模型训练方面却快得多。

**4、C4.5决策树：**

该算法是一个有监督的分类算法，通过信息增益来选择属性和构建决策树。由于受程序工作环境（交换机等网络设备）性能的限制，决策树的层数不可能很高。这样虽然对分类精度有所影响，但却大大加快了模型训练和真实应用的速度。

1. **未来工作**

未来，我团队将继续在徐明伟老师的指导下，与“马上6”公司进行更为深入紧密的联系与合作。对比多种机器学习算法的实现、效果、预测的有效性、准确性。并给予预测的多结果给出网络数据流量优化的方案与思路，争取将本项目从当前的“offline”（离线）分析升级为“online”（在线）分析，并进一步优化各种机器学习算法的性能效果，从而最终实现项目在真实实时网络数据流量中的落地。

1. **参考文献**

[1] Thuy T. T. Nguyen, Grenville Armitage, Member, IEEE, ACM, Philip Branch, and Sebastian Zander,“Timely and Continuous Machine-Learning-Based Classiﬁcation for Interactive IP Trafﬁc”

[2] Jeffrey Erman, Martin Arlitt, Anirban Mahanti , University of Calgary, 2500 University Drive NW, Calgary, AB, Canada,“Trafﬁc Classiﬁcation Using Clustering Algorithms”

[3] Gal Frishman, Yaniv Ben-Itzhak , Oded Margalit , “Cluster-Based Load Balancing for Better Network Security”