**网络流量分类论文阅读报告**

2018210993 毛科龙 深计研18班

我选择的阅读专题为网络安全方面的流量分类专题。

**专题综述**

目前，网络流量分类已经引起了广泛的关注。映射网络流量到其对应的网络应用，在网络安全和管理，例如入侵检测，防御系统，服务质量控制等领域都具有十分重要的意义。网络流量分类尤其在信息安全方面用处非常大。传统的网络流量分类方法主要可以分为两类，基于端口号(port-based)的方法和基于分析数据包形式(datagram format-based)的方法。然而在现代网络环境中，这样的传统方法面临了许多问题。大多数的网络应用，既没有特定的端口号，也没有公开的格式，而且网络中将近百分之五十的流量都是未知流量，甚至有许多僵尸协议，这些僵尸协议本身在设计的时候就没有特定的协议格式。

最近的研究则主要集中于用机器学习的方法来解决网络流量分类问题，主要也分为两种方法：基于包负载的方法和基于网络流的统计信息的方法。前者的思路是寻找出包负载中的协议格式，后者则是利用网络流的统计信息来进行分类。当然，也有一些方法将这两种结合来用。我们可以抽取出包的特征或者流的特征，之后使用机器学习算法，常用的比如SVM，决策树等来对其进行分类。对于加密流量，机器学习算法也可以起到很好地分类效果，因为通常包头内容是不加密的，所以如果用基于网络流的统计信息来分类的话，就可以越过“加密”这道阻碍。T同时，由于网络流量数据巨大，训练集很容易得到，因此，对于网络流量分类，机器学习方法可能会有很大的提升空间，也是目前的主流分类方法。

**评论**

我看的文章如下：

Traffic Classification Based on Flow Similarity

Issues and Future Directions in Traffic Classification

Survey of Traffic Classification using Machine Learning

A Semantics-Aware Approach to the Automated Network Protocol Identification

Byte Segment Neural Network for Network Traffic

第一篇文章Traffic Classification Based on Flow Similarity使用流的相似性来进行网络流量分类，由于传统方法存在的弊端，作者提出了一种负载特征提取方法，类似文档分类，把每一个包的负载看作一个文档，同时，构造出了一个大小为65536的词典，并对每个包进行词频统计。这样，每个包的负载，就可以用一个65536维的向量进行表示，向量的每一维就是该位置所代表词的词频。之后把数据包的向量上下拼在一起，就组成了一个数据流矩阵。对于每一种协议的数据流，我们都可以按如上构造出它的数据流矩阵。当新来一条数据流的时候，我们用该数据流和已有的数据流进行相似性比对，计算每个向量之间的余弦相似性，把该数据流归给和它相似性大的应用协议。这篇文章的做法不仅利用文档分类的方法来对数据包的负载信息做向量化分析，同时也利用了数据包到来的前后顺序，相当于利用了流的一些信息，使得分类的鲁棒性更高。也为后来的研究开辟了新的思路。

第二篇文章Issues and Future Directions in Traffic Classification和第三篇文章Survey of Traffic Classification using Machine Learning 都是综述性的文章。作者详细地介绍了目前网络流量分类方法的发展情况，分析了基于端口号，基于流，基于同一流中包的相关性，基于包负载等方法的优点和缺点以及目前的最好模型(state-of-art)，提出了一些流量分类的现实问题和研究中面临的问题。比如，对于基础的互联网架构来说，现代流量分类任务的复杂性已经超越了最初设计者的想象。由于数据包内容的敏感性，会涉及到隐私的问题，因此在网络流量分类领域不像计算机视觉，自然语言处理等领域有一些公共的数据集可以供大家验证。这就导致了不能统一评判一个模型的好坏，对流量分类领域的发展起了很大的限制。文章也对流量分类的一些子领域做了介绍，比如移动流量，加密流量分类等。这两篇文章行文都非常流畅，把各种方法列成表格，一目了然，为我快速了解网络流量分类领域提供了很大地帮助。

第四篇文章A Semantics-Aware Approach to the Automated Network Protocol Identification提出了一种新的思路。语言是人类交流的工具，类似的，数据包负载中包含的协议内容也是主机之间进行网络通信的工具。因此，作者使用自然语言处理的技术来分析数据包负载进行流量分类。本篇使用了NLP中的LDA主题模型，首先对数据包的原始二进制内容做生成n-gram语言模型的处理。每个数据包的负载内容将由一串n-grams表示，这样，每个数据包的负载就可以看做一篇文档，每个n-gram就是文档中的一个单词。假设每个数据包的负载是由一系列协议关键词生成的，那这里协议关键词就对应LDA中的话题。这样，用LDA模型拟合后，就可以得到每个数据包的协议关键词分布向量，再将该向量作为特征，输入给一些传统的机器学习模型进行分类。该方法完全使用负载信息，因此避免了传输层协议对应用层协议分类的影响。而且这种使用自然语言处理技术来分析包的内容的方法可以适用于文本和二进制负载，对于持续时间长的流量和短的流量都有很好的分类效果。但是也有一些该方法也有一些缺点，因为我们是对每种应用来进行LDA建模，提取该种应用的关键词分布作为特征的，因此对每种应用协议只能做二分类，在实际应用中需要训练很多的二分类模型才能进行多分类，这样对于需要及时分类的场景就不太适用。但是瑕不掩瑜，总的来说，该方法在诸多种类流量的分类上，达到了state-of-art级别，同时这种类比人类交流的思想，也为之后的研究提供了新的思路。

第五篇文章Byte Segment Neural Network for Network Traffic则是使用深度学习的方法。因为RNN对处理序列数据上有很好的效果，所以作者选择以RNN为基础，设计了一种名为BSNN的神经网络来对流量分类。在BSNN中，数据包负载的字节码被分成了许多段，构成输入向量。因为负载的不同段会对分类效果有着不同的影响，因此，引入注意力机制，将向量输入给两级的attention encoders中 ，最终利用focal loss损失函数，通过softmax层输出分类结果。由于近年来深度学习在视觉，语音，文本等领域的突出表现，因此作者尝试使用深度学习方法来处理，避免了人工的特征提取，对许多常见应用取得了很好的分类效果。同时深度学习方法虽然需要的计算资源比较大，但是思路简单，高效，且现在也可以通过预训练技术来缩短时间提升效率，因此很可能替代传统方法和传统机器学习方法，成为未来的主流方法。

**对未来工作的想法**

1. 因为数据包的网络通信可以类比人类语言交流，所以可以关注NLP领域的一些好的方法模型，迁移到网络流量分类的工作中。尝试使用关键词提取，文本分类方面的技术，针对网络流量的情况来设计方法。
2. 深度学习继在计算机视觉领域大热后，在自然语言处理领域也开始逐渐火热，成为不可或缺的方法。因此我们也可以在之前传统分类方法工作成果的基础上，借鉴其精华，设计更好的神经网络来解决流量分类问题。
3. 加密流量、移动流量分类等一些新的方面可能更具有现实意义，因此可以尝试从这些方面切入，寻找高效，可部署的网络流量分类方法。