# 《面向业务的智能网络流量分类方法》调研——使用深度学习的分类模型，按照不同的业务类型，将从节点流出的网络流量进行分类。

## 流量分类的目的、维度、特征等级

### 目的

随着互联网的迅猛发展，越来越多的新型网络应用逐渐兴起，网络规模不断扩大，网络组成也越来越复杂。网络流量分类技术作为增强网络可控性的基础技术之一，不仅可以帮助网络运营商**提供更好的服务质量**，而且能够对网络进行有效的**监督管理，确保网络安全**。

**在本课题中，对网络流量分类的目的主要为区分不同的业务类型，作为题目二中链路选择、流量调度策略的依据。通过流量分类的方法区分应急业务类型，选择合适的链路进行流量转发完成应急业务。**

### 维度

* 应用识别
* **大类识别**
* 用户行为识别

### 特征等级

1）Packet-level的流量分类：

主要关注**数据包**的特征及其到达过程。如数据包的大小分布、到达时间间隔分布等等。

2）**Flow-level的流量分类（最广泛使用）**：

关注**流**的特征及其到达过程。此处的流通常指**[源IP，目的IP，源端口，目的端口，应用协议]组成的五元组**。

3）Stream-level的流量分类：

主要关注主机对及它们之间的应用流量。通常指[源IP，目的IP，应用协议]组成的三元组。适用于粗粒度骨干网络上的长期流量统计特性。

## 流量分类的度量标准\评价指标

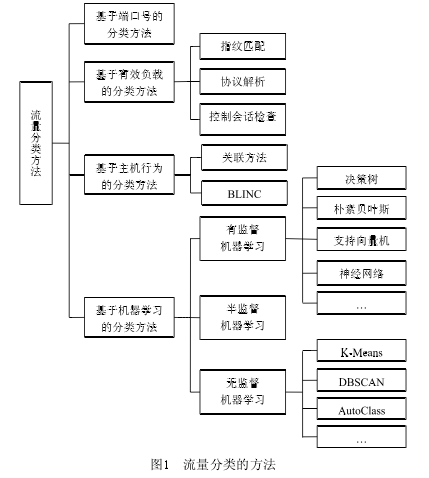
基于Flow-level的流量分类维度，通产用于分类准确率评估标准的有以下几个方面：

* 真正（true positive，TP）
* 假负（false negative，FN）
* 假正（false positive，FP）
* 真负（true negative，TN）

此外，基于机器学习的方法还有其他的度量标准，如

* **召回率（recall）**：recall=TP/(TP+FN)，表示类别A中被正确预测的样本所占比例。
* **精度（precision）**：precision=TP/(TP+FP)，表示在所有被预测为类别A的样本中，真正属于类别A的样本所占比例。

## 流量分类的方法



### 基于端口号的分类方法

基于端口号的方法依赖于对TCP或UDP数据包中端口号的分析，将熟知的端口号进行映射来识别不同的应用类型。位于网络中的分类器只需要找到一次TCP连接中的SYN包，并从这个SYN包中找到目的端口号即可，UDP也使用类似的方法。

这种方法原理简单，适用于高速网络的实时流分类，但是由于新的动态分配端口技术的出现（如P2P）及一些应用使用非熟知端口号进行通信，这种方法的可用性合准确率大大降低。**弃用**

### 基于有效负载的分类方法

为了避免对端口号的过分依赖，提出了基于有效负载的分析方法。该方法通过分析包的有效负载是否包含已知应用的特殊签名进行流分类，具有较高的准确性。

此种方法提高了分类的准确性，也避免了对于端口号的过分依赖。但是分析过程的计算代价大，除此之外，该方法只能识别非加密流量、流量识别过程复杂、无法识别私有协议、且有隐私侵犯和安全性问题。**弃用**

### 基于机器学习的分类方法

基于机器学习的流量分类方法包括两个步骤：分类模型建立 和 分类。首先通过训练集数据建立分类器模型，再使用分类器模型对新数据进行分类验证。在本类方法中，将流量分类抽象为如下几个部分：

* + 网络流类型集合C = { C1，C2，…，Ck}
  + 网络流集合T = { t1，t2，…，tn}
  + 网络流特征向量Ai= ( Ai1，Ai2，…，Aim)

基于此，使用机器学习的流量分类又分为如下几个类别：

1. 基于**有监督机器学习**的流量分类

* 最临近与线性判别式分析法

处理开销大，准确率高

* 贝叶斯算法

贝叶斯神经网络

朴素贝叶斯：实现简单，分类高效。但是依赖先验概率分布，分类效果不稳定。

* 决策树

C4.5决策树法：分类速度快，计算量小，性能有优势。会对样本特征进行过滤，导致忒正选择局部最优，进而导致分类结果不稳定。

* 支持向量机

分类稳定性好，准确率高，而且在小样本训练空间仍保持较好的分类性能。但是在训练集规模大时速度慢。

* 神经网络

1. 基于**无监督机器学习**的流量分类

进行聚合分簇，然后建立各个簇与类的映射关系，在无监督学习中，聚簇类的数量时影响分类算法性能的重要因素。

* 期望最大化算法（EM算法）

AutoClass算法：准确率最高，聚类簇数量最多，但是分类阶段计算代价高

* K-Means算法

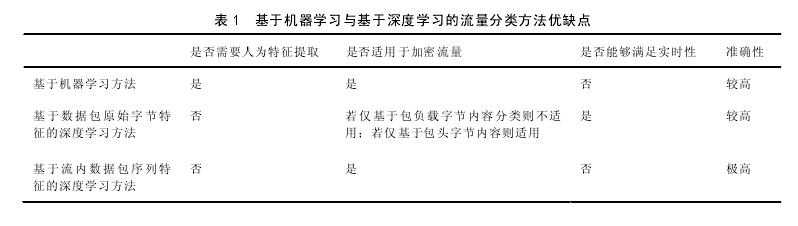
分类准确率随着聚簇类的数量增加稳定上升，但是到达一定阈值以后上升不明显

* DBSCAN算法

分类准确率最低，且聚类簇的数量最少

基于机器学习的分类方法也存在一些问题：机器学习需要专家经验对流量的特征进行提取和筛选，还需要对大量的数据进行标签标注。鉴于此，**基于深度学习的分类方法，可以通过训练自主提取流量特征**，还可以直接基于原始流量数据进行分类。

### 基于深度学习的分类方法



基于机器学习的方法又分为如下的类别：

* **基于数据包的原始字节特征——输入数据包原始的字节内容**.

基于数据包原始字节特征的深度学习方法的优点是可以直接根据数据包的字节内容做推理，做到实时分类，但分类结果依赖于数据包负载的内容。当流量加密时，数据包负载部分不再可用；而数据包的 IP 地址和端口字段也会对分类效果造成极大的影响，带来**过拟合**问题

* **基于流内数据包序列特征——输入流内数据包的大小、包时间间隔序列等特征**

基于流内数据包序列特征的深度学习方法的优点是不依赖于数据包原始内容，对于加密流量更灵活，但需要等待一段时间的数据包以构成时序序列，因此分类**实时性差**。

## 挑战

**数据集**、部署开销、**实时流量分类**、新型应用的识别、精细化流量分类（特定协议上承载应用类型，特定应用不同功能模块等）、加密流量分类等等。

## 下一步

1. 计划使用深度学习的分类模型实现流量分类。
2. 学习机器学习理论和方法。调研、复现相关的论文，找数据集

小论文发什么水平？

提高分类准确率，能否第一个点发小论文？

结合迁移学习，从其他模型中提取特征参数，用迁移学习的方法训练网络流的数据集，若实验则可证明可行，则在分类问题中，只要采集少量数据集就可以进行分类器训练。？

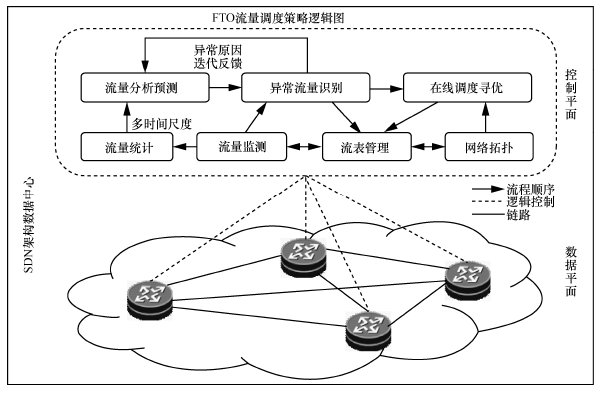
考虑一个新方法，提高DL流量分类的实时性及准确性，以适用于应急场景下？

# 基于业务分级与链路状态的智能流量调度策略

**经过题目一，将网络流量按照业务进行分类，由于不同的应急业务优先级不同，且应急场景下的链路状态也不尽相同，存在带宽、时延等方面的差异。课题二的目标就是结合业务等级和链路的状态信息，为业务选择最优传输链路。**

## 调研论文：基于斐波那契树优化算法的数据中心流量调度策略（王耀民，2019）

以运营数据中心为研究对象，同时考虑网络结构和网络流量两方面因素，设计一种 SDN 架构数据中心的流量分析预测在线调度机制，提出基于斐波那契树优化（FTO,  Fibo-nacci tree optimization）算法的数据中心流量调度策略（下文简称 FTO 流量调度策略）



## 调研论文：基于流量特征的流调度策略研究综述（2020）

在网络传输过程中，根据网络流量的特征，对数据流进行有区别的调度，这对网络资源的高效利用具有重大意义。本文针对现阶段大流和小流的调度策略进行了综述，并对流调度工作中一些亟需解决的问题进行了讨论。

流调度是以**最小化流完成时间、最小化数据包传输的错误率和丢包率、最小化延时、最大化网络资源利用率和公平性为目标的**。合理的流调度方案可以优化网络传输，避免网络拥塞，实现网络的负载均衡和网络资源的合理利用，对于数据中心网络性能的提升，优化用户体验有重要意义

### 基于SDN的数据中心网络

在数据中心网络中融入 SDN，形成软件定义—数据中心网络( software defined data center network，SD-DCN) 。SDN的可编程能力，使数据中心可以灵活地制定控制策略，避免了大量的命令行操作，减少了网络管控的复杂度。

数据中心网络中的流量**表现为明显的大小流特征**，大象流和老鼠流的数据大小和性质不同，并且它们也有着不同的服务要求。**大象流对网络带宽要求高，老鼠流则对时延较为敏感。**

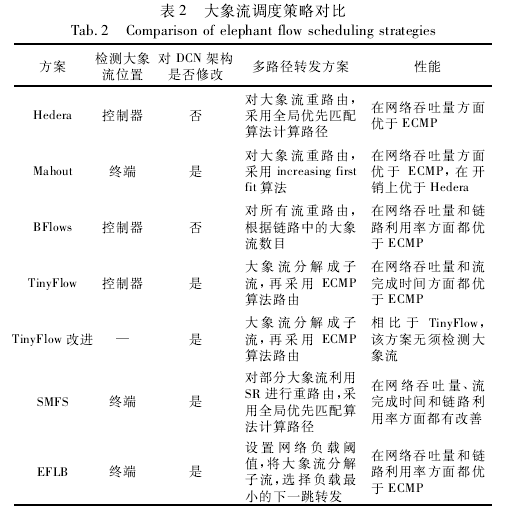
传统IP网络通常使用最短路径路由算法（OSPF等）进行数据转发，没有考虑路径实时状态——导致拥塞和浪费.

最短路径优先算法SPF、基于哈希的等价多路径路由ECMP——未考虑链路负载和流大小

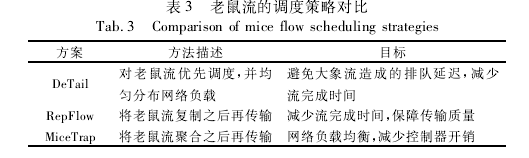
### 基于流量特征的流调度策略

1. 大象流的调度策略

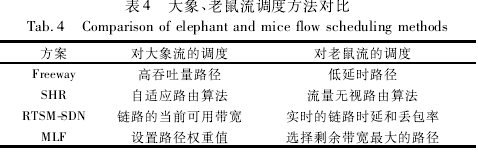
* 利用 SDN 全局网络视图和动态规则配置能力，对大象流进行动态调度，选择无冲突路径，避免大流冲突;
* 将大象流分解成多个子流，多路径传输



1. 老鼠流的调度策略



1. **大象流和老鼠流的调度策略（比较贴近毕业论文的应用场景）**



### 研究展望

* 自适应地提取和识别流量特征。.
* 分布式网络环境下的流量特征。
* 实现多个优化目标的流调度。
* **网络智能化。**知识定义网络( KDN) 是一种依赖于 SDN和机器学习( ML) 运行以优化网络性能的概念。将机器学习的方法应用到流量调度中。——

下一步：

第二个点 类似于路由选择算法吗？

论文中的应用场景是供应商数据中心数据流量很大的情况，但是应急场景下是否符合这个要求？

下一步调研的重点：路由选择算法/网络智能化/应用场景？