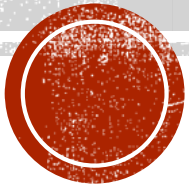
网络流量分类进展



前一段工作总结

- 机器学习单分类器实验
- 调研网络流量分类方法
- 从原始包数据中分流
- 流特征提取
- 流标记

整体方案

- 流量数据获取（√）
- 解析获取信息（√）
- IP地址匿名化（√）
- 调研不同分类方法（√）
- 包特征提取（√）
- 基于端口号进行流量分类（√）
- 基于半监督机器学习流量分类（doing）
- 总结实验结果到论文中

机器学习单分类器实验结果

- 基于端口号的分类实验结果
- 根据网络流量分类的特点采用了几个单一分类器
- 准确率极高，大部分能达到99%左右的准确率

```
label set num: 10000
KNN Accuracy: 0.993
Logistic Regression Accuracy: 0.9505
Random Forest Accuracy: 0.9975
Decsion Tree Accuracy: 0.9975
GBDT Accuracy: 0.9975
AdaBoost Accuracy: 0.9635
GaussianNB Accuracy: 0.997
Multinomial Naive Bayes Accuracy: 0.477
```

网络流量分类方法

- 经过前期调研，现在我们有了如下的大致思路
- 总体上采用半监督机器学习
- 首先利用基于端口号的分类方法对流进行标记
- 然后对未标记的网络流量使用聚类算法进行分类识别
- 对比“仅采用流的前几个包来分类”的效果（二者效果现在还不知道）
- 最后，将应用归类，一类是时延敏感，一类是带宽敏感，其他的归为不确定

从原始包数据中分流

- 基于三元组分流（目前采用的方案）
- 由于我们获取的数据经过NAT处理，故而流定义中的四元组中的源IP地址不可用
- 因此我们采用三元组（源端口、目的IP、目的端口）来区分一条流

- 基于握手（已经实现，但不采用）
- 三元组+握手syn/fin信息
- 实验效果不太好，分出的流中数据包太少，且分出的流也很少

流特征提取

- 包特征提取
- 端口号
- 时间戳
- 包长度
- 有效负载长度

- 流特征提取
- 包总数
- 连接的持续时间
- 包的平均长度
- 包之间平均间隔时间

流标记

- 根据端口号对每一条流的应用类型进行标记
- 很多端口号是没有应用类型的，标记为”unknown”
- 未来将增加“时间敏感”/“延迟敏感”的标记

THANKS