## **3** 匿名流量检测系统

3.1 匿名流量检测流程

本实验首先对数据集进行了特征提取、数据预处理等操作，然后使用改进的深度森林模型从混合流量中检测出 Tor 流量如图3.1所示。

图3.1Tor流量检测流程

3.2 Tor匿名流量特性分析

Tor原始流文件中提取的数据集中包含大量的空值高纬稀疏数据，严重影响对Tor流量预测的准确性。本文通过分析Tor流量的特性，通过筛选用较少的特征便可以精准描述Tor流量行为。

用户使用Tor客户端发送请求，在Tor匿名通信系统中默认需要经过是三个节点的转发才能到达目标服务器，而且三个节点一般部署在不同国家的不同地区，因此Tor流量的转发时延相对于正常流量会大，可以使用数据包时延、流持续时间等特征描述转发时延差异。

通过研究发现Tor一般将应用程序数据包封装成固定的长度，所以Tor数据包长度的变化范围是有限的，Tor和NonTor流量在数据包长度特征上存在明显差异，可以用此特征来描述数据包大小变化的情况和差异性。

通过研究发现用户在建立Tor连接之后，如果在一段时间内应用程序不进行任何操作，网络处于空闲状态；如果应用程序进行相关操作，将会产生持续的网络访问，网络处于忙碌状态，Tor匿名通信系统无法隐藏这些网络流特征，因此可以使用流变成活动状态之前的空闲时间和流变成活动状态的空闲时间的统计特征描述Tor这一特点。

综上所述， Tor 虽然通过加密和封装等方式将非正常流量隐藏于普通流量，但是Tor匿名网络具有不可避免的时延、流持续时间、数据包长度分布、数据包时延等特性仍然存在，根据对这些特征的处理，依然可以区分Tor和NonTor流量。

3.3 Tor匿名流量识别特征提取

为了降低原始数据集中特征的冗余度，提高检测的准确率，本文在Tor匿名流量和NonTor匿名流量数据集上对CIC Flow Meter产生的83种流量特征进行了评估。

本文尝试通过提取出每个流超时时间的数据集的基于时间的特征，我们将Tor数据集中的所有流标记为Tor。将其和相同流超时时间的加密NonTor数据集进行合并用作输入。对每个流超时时间数据集（2s，4s，8s，16s，32s，64s，128s）应用多种不同的特征选择方法，最终我们选择使用Infogain+Ranker特征选择算法处理16s数据集得特征取得了最佳的实验结果如表3.1所示。

表3.1 特征选择结果

|  |  |
| --- | --- |
| 特征 | 相关度 |
| Flow Bytes/s | 0.8778 |
| Flow Duration | 0.8745 |
| Fwd IAT Mean | 0.7135 |
| Bwd IAT Mean | 0.7134 |
| Flow IAT Std | 0.7034 |
| Fwd IAT Min | 0.6961 |
| Flow IAT Mean | 0.6921 |
| Bwd IAT Min | 0.6874 |
| Idle Max | 0.6868 |
| Idle Min | 0.6755 |
| Idle Mean | 0.6726 |
| Flow Bytes/s | 0.6692 |
| … | … |

如上表所示，本文最终选择了特征相关度前35的特征作为Tor匿名流量的时间相关性特征它们具有Tor匿名流量识别的最大信息增益和卡方统计量。数据集提取的特征属性及其描述如下表3.2所示。

|  |  |
| --- | --- |
| 特征属性 | 描述 |
| Fwd IAT mean/max/min/std | 正向流中包到达间隔的 mean/max/min/std |
| Fwd Packet Length mean/max/min/std | 正向流中包长度的 mean/max/min/std |
| Bwd IAT mean/max/min/std | 反向流中包到达间隔的 mean/max/min/std |
| Bwd Packet Length mean/max/min/std | 反向流中包长度的 mean/max/min/std |
| Flow IAT mean/max/min/std | 包到达间隔时间的 mean/max/min/std |
| Active mean/max/min/std | 流变为空闲态之前活动时间的 mean/max/min/std |
| Idle mean/max/min/std | 流变为活动态之前空闲时间的 mean/max/min/std |
| Total Fwd Packet | 正向数据包总数 |
| Total Bwd packets | 反向数据包总数 |
| Total Length of Fwd Packet | 正向数据包总大小 |
| Total Length of Bwd Packet | 反向数据包总大小 |
| Flow Bytes/s | 每秒流字节数 |
| Flow Packets/s | 每秒流数据包数 |
| Flow Duration | 流持续时间 |

3.4 流量数据预处理方法

为了获得更加有效的Tor和NonTor流量数据，本文对其进行预处理。

3.4.1 消除重复数据流

通过CIC Flow Meter设定流超时时间将不同类型的pcap文件保存到csv文件中并将其合并，通过研究发现无论是Tor数据集还是加密的NonTor数据集都存在大量的重复数据流，并发现有的重复率高达30%。这些不仅但没有研究价值，反而容易使得实验结果虚高。

3.4.2 数据归一化处理

通过无论是Tor数据集还是NonTor数据集其中特征的取值范围跨度较大，例如Flow IAI Max的取值范围为，这将严重影响模型的收敛速度和分类性能。于是本文采用Min\_MAX数据归一化方法，将所有特征范围归一化到之内。

3.5 检测目标定义和效果评价方法

Tor流量行为检测的主要目标就是从混合流量中检测Tor匿名流量，属于二分类实验。

在匿名流量识别实验中，数据集中样本往往都是不平衡的。若正负样本比例差别非常大的情况下仍依靠准确率评判是不科学的。例如正常样本有99000条数据，异常样本只有1000条数据。在识别过程中即使所有的数据全判断为正常样本准确率也可以高达99%，此时准确率指标不仅不能表示性能很好，相反此结果是非常差的。本文使用准确率（Accuracy）、精确率（Precision）、召回率（Recall）和F1-Score作为评价指标来综合判断Tor匿名流量识别的效果。

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| DecisionTree | 99.69 | 99.07 | 98.99 | 98.9 |
| CNN |  |  |  |  |