**加密的恶意网络流量分类**

**郭嘉 2111116**

**一、背景**

网络空间在生产生活中扮演着越来越重要的作用，恶意网络攻击对于国家、政府、企业、个人而言都是巨大的威胁，因此设计快速检测出网络流量是否为攻击者发出的算法是非常有意义的。然而现在越来越多的网络流量通过TLS、IPsec等技术加密，这为防御者防御恶意流量带来了巨大的挑战。然而加密的恶意攻击流量具有一定的流量模式特征，可以通过机器学习算法[3]检测其特点，然而[3]采用了许多可解释性较差的模型，如随机森林、Xgboost，集成学习尽管在准确性等模型表现上有较好地表现，但是他们的可解释性仅仅来源于特征重要性，这对防御设施设计者来说是不足的，简单而有效的过滤条件将减轻防火墙硬件和软件的压力，并且他们对不同类型的攻击也没有做出非常详尽的分析。因此本文希望设计兼顾解释性和准确性的模型，并尽可能的在加密场景下分析出不同攻击具有的特点。

**二、难点**

**1．加密流量**

加密流量分析要求我们不利用报文信息就鉴别出某流量是否为恶意流量，这似乎是非常困难的事情，但其实攻击有其特定的流量模式，有些攻击短而紧凑，有些攻击发送大量小包，有些攻击发送少量大包，这并不需要我们知道加密流量内容，即便如此，如何有效地利用这些流量特征进行判断也并非易事。

**2．攻击类型判断**

有时仅仅判断流量是否是异常或具有恶意并不能满足我们的要求，攻击类型的判断能帮助我们更好地做出针对性防御措施，然而攻击类型的判断需要我们对恶意攻击的攻击模式有更深入的认识。

**3．可解释性与模型精确性平衡**

有许多机器学习模型都能很好的利用所给特征进行分类，RNN、CNN等深度学习模型精度可能很高，但他们的可解释性非常差，这并不能给防火墙或者防御措施设计人员带来任何有用的信息，唯一能指望的是把他们直接塞进防火墙软件，但高昂的计算资源需求使人望而却步。我们认为网络攻击的模式都十分有特点，利用简单的分类器如Logistic回归和决策树就能达到不错的效果，而且其模型训练出来后还能给防御措施设计者带来许多有启发的信息。

**三、技术方案**

**1．数据集选取**

尽管我们希望解决加密恶意流量分类的问题，那么是否就意味着我们要完全放弃掉不使用加密协议的网络流量？我们认为，只要不利用报文内容进行分析，而仅仅通过流量特征分析恶意性的方法就能称为加密恶意流量分类方法。基于这样的考虑，本文使用以下两个数据集[1]CSE-CIC-IDS2018与[2]CTU-13。

[1]CSE-CIC-IDS2018，通过50台主机对指定的520台主机、30个服务器发动了7种类型的攻击，攻击类型包括暴力破解攻击、Heartbleed攻击、僵尸网络攻击、DoS攻击、DDoS攻击、网页攻击、内部渗透攻击，每种攻击类型都通过某几种特定技术手段实施。他们利用CICFlowMeter软件从收集到的pcap文件中提取特征，共得到80种特征。该数据集可以在Kaggle上下载https://www.kaggle.com/solarmainframe/ids-intrusion-csv，由于我们只分析特定的DDoS攻击、DoS攻击以及暴力攻击，因此只下载02-14-2018.csv、02-15-2018.csv、02-21-2018.csv三个文件。

[2]CTU-13是一个具有多类型非模拟僵尸网络攻击、未知网络流量掺杂、多种僵尸网络同时存在等特征的僵尸网络检测数据集。他们利用tcpdump在Linux主机上收集被特定恶意程序感染的虚拟机发送的流量并在学校路由器上搜集所有网络的流量，最后用Argus软件把pcap文件进行预处理生成NetFlow文件。数据集共有7种类型的僵尸网络，13个场景，NetFlow文件中每条数据通过Start Time, End Time, Duration, Source IP address, Source Port, Direction, Destination IP address, Destination Port, State, SToS, Total Packets 以及Total Bytes属性描述，并可以从标签中提取出其是否为恶意流量以及其属于哪一类僵尸网络产生的流量。该数据集可在https://www.stratosphereips.org/datasets-ctu13下载，我们只取13个文件夹中每个文件夹下的NetFlow文件进行处理，我们不考虑pcap文件和exe可执行文件。

**2．任务**

**1） 攻击鉴别。**我们希望将某种类型的攻击与正常网络流量区分开来，暴力攻击与正常流量的分类在02-14-2018.csv数据上完成；DoS攻击与正常流量的分类在02-15-2018.csv 数据上完成，DDoS与正常流量的分类在02-21-2018.csv上完成。即我们希望完成三个二分类任务。

**2）攻击技术分类。**FTP暴力攻击与SSH暴力攻击是两种不同的暴力攻击类别，我们希望通过有效的机器学习算法将他们快速区分开来；同样我们希望区分DoS攻击使用GoldenEye技术还是Slowloris技术以及DDoS攻击使用HOIC技术还是LOIC技术。我们设置三个类别，如对于暴力攻击的分类，我们设置FTP、SSH、normal共三类，我们希望完成三个三分类任务。

**3）僵尸网络流量鉴别。**该任务在CTU-13数据集上完成，目的是通过仅有的十来个流量特征判断流量是否来源于僵尸网络，此为二分类任务。

**4）僵尸网络分类。**该任务在CTU-13数据集上完成，目的是通过仅有的十来个流量特征判断流量来源于7种僵尸网络：Neris、Rbot、Virut、Menti、Sogou、Murlo、NSIS,ay（后文简称NSIS）的哪一种，我们设置8个类别，最后一个类别为正常流量。数据筛选与2.3中所述一致。

为什么除了鉴别流量是否具有恶意外，我们还需要鉴别该恶意流量的具体类别？以下从两个角度分析。首先，恶意流量具体的类别标签大大减轻了机器学习的学习压力，监督学习往往比无监督学习表现更好，具有丰富标签或精细标签的监督学习比仅具有少量标签的监督学习表现更好是人们的共识。即便在实际测试时仅需要区分是否为恶意流量，仅需把所有类型的恶意流量看成一类即可。其次，具体的类别信息有利于机器学习模型给出更具判别力的参数，如后文提到的决策树为每类攻击指定了具体的判别条件，利用这些条件防火墙等防御设施能够更有针对性的防御特定类别的攻击。

**3．特征选取**

对于CSE-CIC-IDS2018数据集，我们并没有对原始数据的特征进行处理，因为他们已经通过CICFlowMeter软件提取了80个十分有用的流量特征，具体的特征信息可以在https://www.unb.ca/cic/datasets/ids-2018.html中看到。对于CTU-13数据集，我们把原始数据集中的Start Time, End Time，以及地址端口等信息去除方向特征Dir和状态State也不予以考虑。我们也没有像[3]中一样使用软件提取conn.log,ssl.log以及x509.log的信息，这是因为我们认为，如果报文不仅仅在应用层上加密，而且在传输层上也作了加密，那么这些特征将会很难获得，一个典型的例子就是运行在网络层上的加密IPsec。

**4．用逻辑斯蒂回归直观分析各特征影响**

**1）模型介绍**

逻辑斯蒂回归模型常用于二分类任务，给定个网络流量特征 预测流量为恶意的概率如下式所示，其中为模型参数。

（1）

当预测概率大于0.5时则判断为恶意流量，通过简单的运算可以得到当流量特征通过加权后满足如下条件时：

（2）

判定为恶意流量，且该加权和越大，模型对该流量为恶意流量的肯定性越大。模型参数为使交叉熵损失最小的参数，有许多优化算法能求得这些参数，而我们使用拟牛顿法。对于多分类任务常有One Vs One和One Vs Rest的方法，出于可解释性和时间复杂度的考虑，我们选取后者。

**2）利用逻辑斯蒂回归模型解释恶意流量特点**

除了精确地分出类别，我们更希望模型能告诉我们一些关于恶意流量特征的有用信息。图1为利用CTU-13数据进行训练僵尸网络鉴别模型所得到的logistic模型所能带给我们的信息。为了理解这张图，首先注意到2式，特征加权和越大，判别为恶意流量的概率也就越大，这就说明如果训练出的为正值，则说明如果我们碰见大的，就应该更加确信该流量为恶意流量。这里我们对协议类型特征采取了One-hot编码，如果某条流量信息对应igmp协议，则igmp对应特征设为1，其它协议对应特征设为0。我们对的绝对值进行排序，得到中间一列数值，他们对应的特征名字在左侧一列。先看第一行，igmp对应的重要性为

-10.97，这说明如果某条流量信息使用的是igmp协议，则很有可能该流量为正常流量，由于-10.97对应的特征权值是所有特征权值里绝对值最大的，所以我们知道某流量是否使用igmp协议是判断该流量是否为恶意流量的重要依据。同样我们可以看出，如果源主机发送信息的速率很大，那么其为恶意流量的概率也相应较小；如果某流量使用tcp或者rtp协议，其可能为僵尸网络流量的可能性相应也会增大。

细心的读者应该了解到如果某个特征普遍比其他特征大几个量级，那它的加权值也不会太大，事实上，我们对特征先进行了归一化预处理，避免了这种情况，因此加权值确实能够很好反映他们对应特征在僵尸网络流量鉴别上的重要程度。

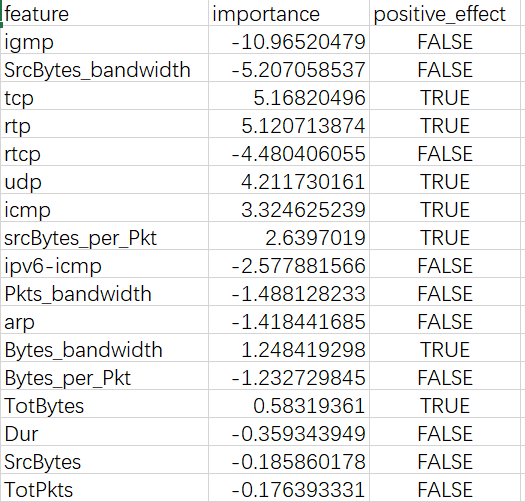


图1

**3）为什么使用logistic而不是带核SVM**

从2式可以看出，logistic回归和线性核SVM思想很类似，都是希望通过超平面分割不同类别样本，然而现在面对CTU-13数据集的NetFlow文件，我们仅有十几个看起来不那么具有判别力的特征，我们很有可能处于一个低维特征情形下，为什么不用带核的SVM将特征映射到高维空间使得它们线性可分呢？主要是我们认为这种方式会使得模型的可解释性降低，而可解释性是设计网络防御策略中对设计者最具启发的一步，我们不希望把模型丢给防火墙让其跑一些复杂的模型，我们希望防火墙能够凭借一些简单的过滤条件就能区分出恶意流量。而且我们推测不论恶意攻击如何发展，其都有非常鲜明的流量特征，因此理论上都应该能够通过简单的过滤条件将其过滤掉，而不是借助包含复杂非线性映射的机器学习模型。

**4. 用决策树从繁杂特征中挑出最重要的**

**1）模型介绍**

如果不谈及训练，仅谈及利用模型进行推断，那么决策树模型其实不用作过多介绍，即便对该方面不了解的读者看到图2也能了解到决策树是如何帮助我们作出分类的。在图2对应的场景下，我们希望利用网络流量特征对DDoS攻击的类型进行分类。从该决策树可以看出，当通信过程中两个流量的最小间隔时间小于1.5个时间单位，我们就能确定该次通信为DDoS\_LOIC攻击建立的，否则我们认为该通信要么为正常通信，要么为DDoS\_HOIC攻击建立的。此时再分析子流发送包的数量，如果发送数量为4及4以下，我们能够有非常大的把握认为该通信为DDoS\_HOIC攻击所建立的。

以上我们对决策树的推断已经做了详细的叙述，那么如何从CSE-CIC-IDS2018数据集中80个特征中找到这些具有强大判别力的特征呢？这其实也是决策树训练的基本思想的来源，我们只需要比较这80个特征，哪个特征有能力把根节点的数据分成纯净度较高的两份，这里纯净度的意思指的是尽量属于一个类别，一般可以采取gini指数或entropy来衡量该纯净度。这里的有能力指的是存在一个阈值，这个阈值能较好地分划数据集。

**2）为什么有逻辑斯蒂回归还要用决策树**

由于CSE-CIC-IDS2018中给了80个流量特征，对于这么多可用的特征，逻辑斯蒂回归有很多种设置权重的方式使得模型在训练集上表现良好，而在测试集上因为过拟合而表现不佳。还有一个重要的原因是，决策树能直接从80个流量特征里面找到最具有判别力的几个特征，而避免考虑其他一些无关的特征，这从图2中就能直接看出来，我们从80个特征中选出了两个对判断DDoS攻击最具判别力的特征，而防火墙只需要根据这两个信息设置过滤条件即可保证过滤掉LOIC和HOIC引发的DDoS攻击。这种仅根据少数特征进行判断的方法对减轻防火墙信息处理压力，以及从庞杂的网络流量中快速过滤掉有害信息是至关重要的。

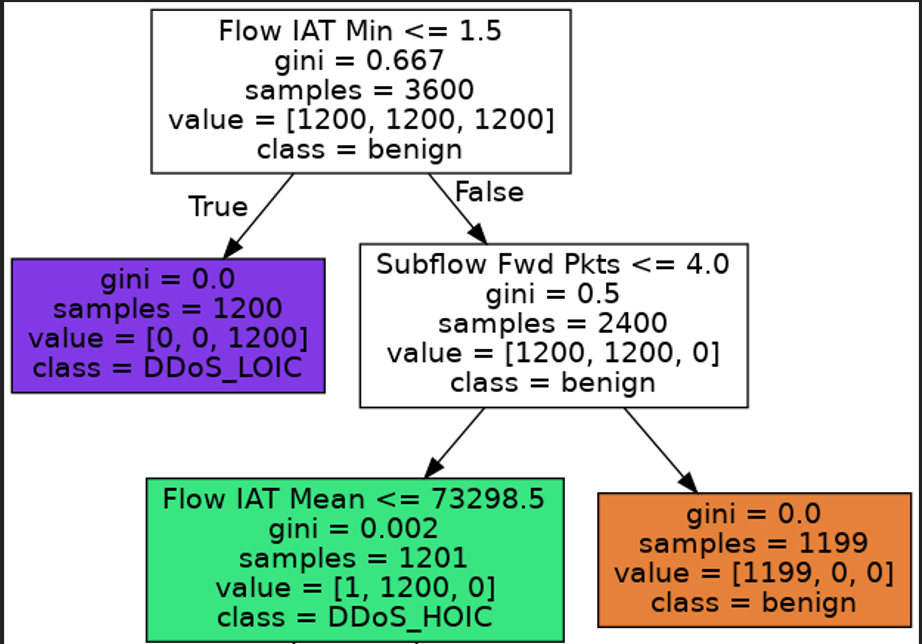


图2

**5. 通过梯度提升树为防御者提供多个有用视角**

**1）模型介绍**

这里不对算法作详尽描述，只阐述主要思想。利用梯度提升树进行恶意流量分类的主要思想是利用多棵树预测一个与恶意概率正相关的值，求损失在此预测值处的梯度，通过构造新的树预测负梯度，并把新的树加入到原来树的集合中，通过这样不断迭代产生一系列新的树，直到梯度值小于某个预定阈值为止。本质上就是梯度下降的思想，只不过梯度下降是利用添加不同的新树完成。

图三是在CTU-13数据集上用来预测僵尸网络流量所有梯度提升树中的其中一棵，我们提前把最大深度设置为4。可以看到到每个叶子节点的最下面一行对应了一个value，大于0我们就认为该流量为僵尸流量，这个value值越大则说明我们越有把握说该流量为来自僵尸网络的流量。可以看到梯度提升树与logistc回归给我们带来了不同的视角，这里先通过Pkts\_bandwidth即发包速率确定第一步的分划标准，发包速率大于135.19个每秒则很有可能是来源于僵尸网络的流量。但观察最后一行第五个染白的节点，如果与此同时有源主机每个包所包含字节数较小，且所有包的每包字节数较小，则非常有可能该流量为正常流量。当然，最后的预测除了参考该树，还需要将其他树的预测结果通过预先设定的权值加和得到。

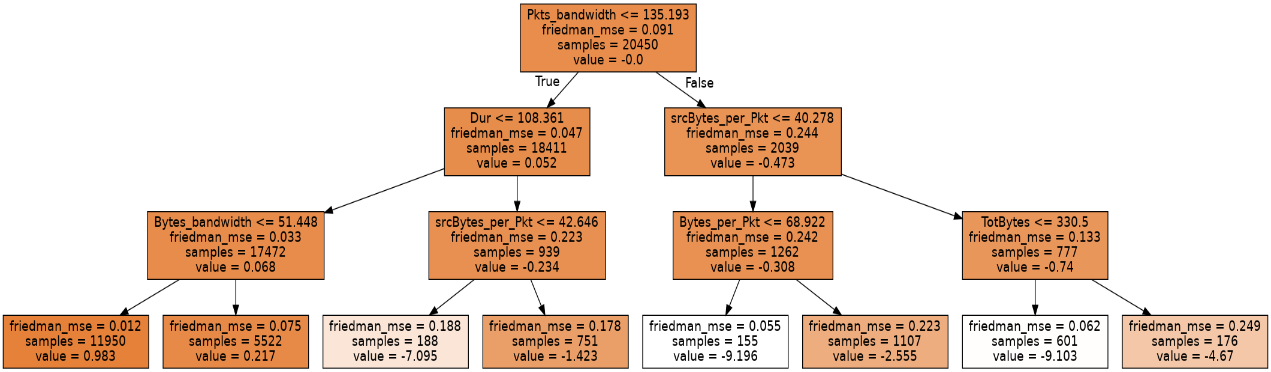


图3

**2）为什么有决策树还要用梯度提升树**

从上节例子可以看出，梯度提升树非常直观的通过value的值判断了来源于僵尸网络的概率，我们不仅需要知道如何分类，还需要知道通过某个分类依据对预测结果的把握有多大，如图3最后一行第五个染白的节点，value的绝对值达到9.196，比其他节点都要大的多，这说明其父节点Bytes per Pkt特征对区分僵尸流量具有十分重要的作用，而决策树仅仅预测了一个类别信息，尽管从gini指数等指标可以看出一个节点的重要性，但是gini指数与实际预测概率之间的关系并不直观。

同时，梯度提升树通过不同树的加权和来确定最终的结果，每棵树都能对僵尸网络流量的特征提供新的视角，而单一的决策树很难完成这个任务。

那么我们什么时候会放弃掉梯度提升树呢？在本文中提到的DDoS攻击方式分类，DoS攻击方式分类，以及暴力攻击方式分类的情形下，我们放弃了梯度提升树。其实从图2可以看出，在DDoS分类的任务中，我们通过决策树已经找到了那些最具有判别力的依据，达到了很高的精确度：99%以上。然而在僵尸网络分类的情形下，利用决策树只能达到84%的分类精度。这时单一决策树能提供信息或许太少，我们就要考虑从多棵决策树上找到僵尸网络的流量特征。尽管梯度提升树在该任务上准确度只有79%，但是从多个视角分析僵尸网络对后续防御者的进一步分析十分关键，精度并不是判断模型好坏的唯一指标，还需看到模型为防御者的后续分析提供了多少有用的，有启发的信息，从这点来说梯度提升树是一个很好的分析手段。

**四、实验**

**1. 数据选取**

对于CSE-CIC-IDS2018数据集，我们分别在02-14-2018.csv、02-15-2018.csv、02-21-2018.csv三个文件上独立进行实验。对于DDoS攻击分类和鉴别任务，我们随机选取其中1500条HOIC攻击流量，1500条LOIC攻击流量数据，1500条正常。对于DoS攻击分类和鉴别任务，我们随机选取5000条GoldenEye攻击流量，5000条SlowLoris攻击流量，5000条正常流量。对于暴力破解攻击，我们选取5000条FTP暴力破解流量，5000条SSH暴力破解流量，5000条正常流量。

对于CTU-13数据集上的僵尸网络分类和鉴别任务，我们从所有NetFlow数据中随机选取2000条正常流量，从13个场景中每个场景随机选取至多2000条恶意流量。

以上两个数据集的每个实验中，我们都采取4:1的比例分割训练集和测试集。

**2. 模型衡量标准**

由于在某些任务上存在类别不平衡问题，如DDoS攻击鉴别中攻击样本和正常样本的比例为2：1，此时即使将所有样本分类为攻击也能达到66.7%的准确率，因此为了较好地衡量模型在类别不平衡的准确率表现，我们采取平均召回率AR作为准确率衡量标准，并对鉴别任务计算AUC。

**3. 模型使用**

对于CSE-CIC-IDS2018数据集的几个任务，我们均只采用决策树算法，这是因为我们发现只使用很浅的决策树就能在这几个任务上取得非常好的表现，并且能非常好的解释这些攻击的特点，此时没必要用其他算法进行攻击特性分析。

对于CTU-13数据集，该数据集数据不平衡特点很突出，Sogou攻击只有十来个样本，而且特征较少，这造成使用这些数据分析较为困难，因此我们采取多种分类方式进行实验，目的是从多角度解释这些僵尸网络的特性，利用不同模型的优点进行尽善尽美的分析，包括Logistic回归，决策树和梯度提升树。在前文已对梯度提升树进行僵尸网络流量分类的实验作过一定分析，由于其时间复杂度较高，不适于用在多分类，因此实验部分不再对其进行说明。

**4. CSE-CIC-IDS2018攻击分析**

**1）DDoS攻击分析**

我们在2-5中选取了几个深度进行了实验，最后在权衡精确度与可解释性的考虑下，选取了2作为最大深度。与从图4可以看出，如果源主机发送的包平均大小小于154.3 我们能够有非常大的把握认为其为DDoS攻击产生的流量，较为次要的依据是每秒发包数，每秒发包数如果大于29.77我们就能很肯定地说这流量来源于DDoS攻击。图10左上为DDoS攻击检测的ROC曲线，对应的AUC为99.997%，AR为99.997%，攻击分类的AR为99.998%。

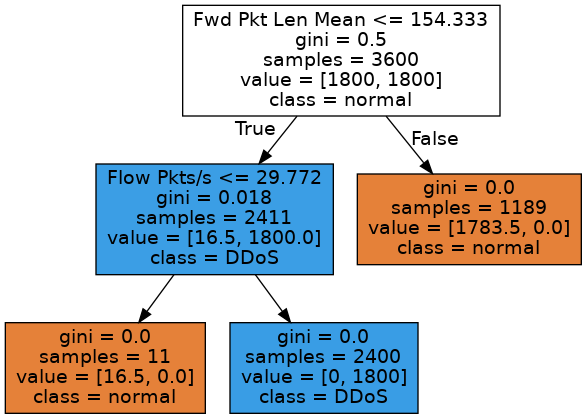


图4

**2）DoS攻击分析**

我们在2-5中选取了几个深度进行了实验，最后在权衡精确度与可解释性的考虑下，选取了3作为最大深度。从图5可以看出，划分正常流量和DoS攻击的主要依据是原主机发送的最小片段大小，如果最小片段大于30，则很有可能为DoS攻击所产生流量。为了进一步区分DoS攻击方式，需要判断目标主机的返回包平均间隔，如果平均间隔时间小于等于12945202，那么很有可能是GoldenEye攻击，否则，则很大把握是SlowLoris攻击。图10右上为DDoS攻击检测的ROC曲线，对应的AUC为98.90%，AR为98.90%，攻击分类的AR为95.07%。

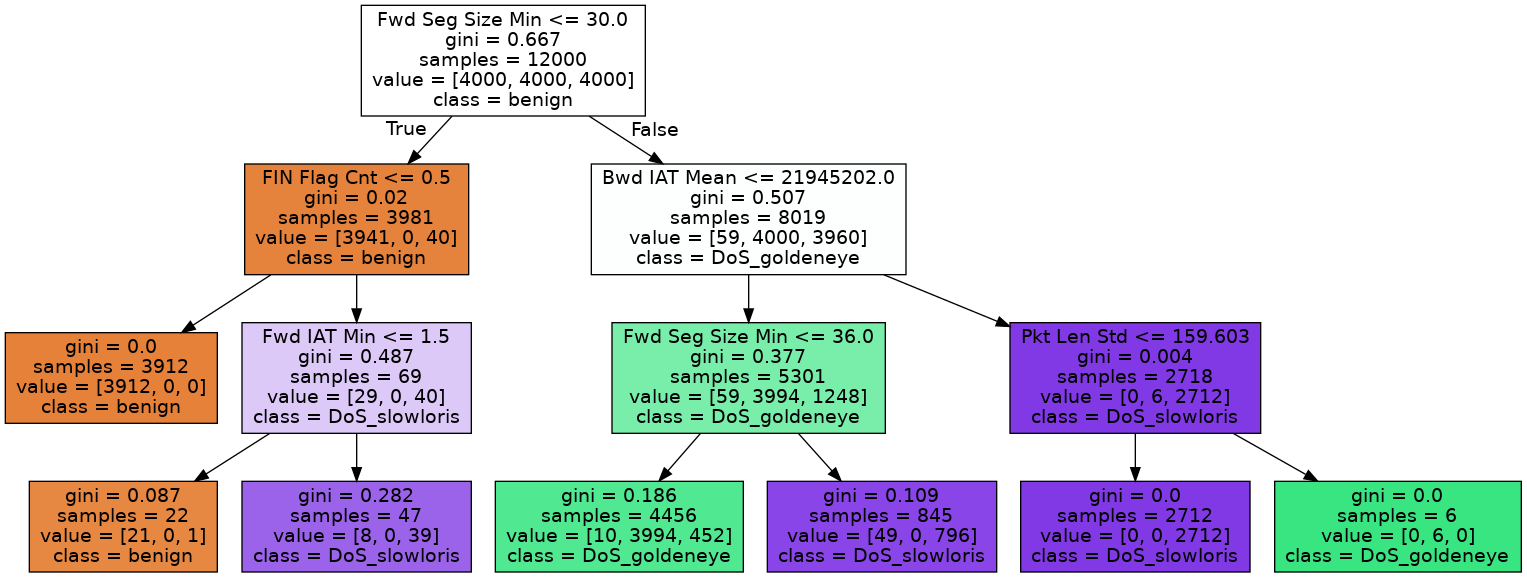


图5

**3）暴力攻击分析**

我们在2-5中选取了几个深度进行了实验，最后在权衡精确度与可解释性的考虑下，选取了2作为最大深度。从图6可以看出，原主机发送的最小片段程度大于36，则可以判断流量来源于FTP暴力攻击。如果原主机发送的最小片段程度小于等于36且大于30，则可以判断流量来源于SSH暴力攻击，其余为正常流量。图10左下为DDoS攻击检测的ROC曲线，对应的AUC为99.95%，AR为99.95%，攻击分类的AR为99.97%。

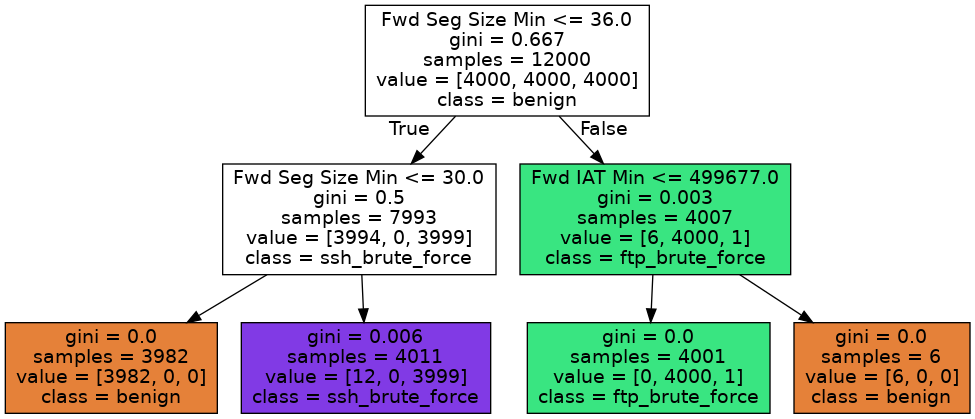


图6

**5. CTU-13攻击分析**

**1）通过logistic混淆矩阵分析不同僵尸网络攻击特点**

图7为logistic回归预测僵尸网络类别的混淆矩阵，可以看出有很多Virut流量被模型认为来自Neris僵尸网络，这说明Virut僵尸网络与Neris有一定的相似之处，也有许多Murlo流量被模型认为来源于Menti僵尸网络，这说明Murlo僵尸网络与Menti僵尸网络也有一定的相似之处。

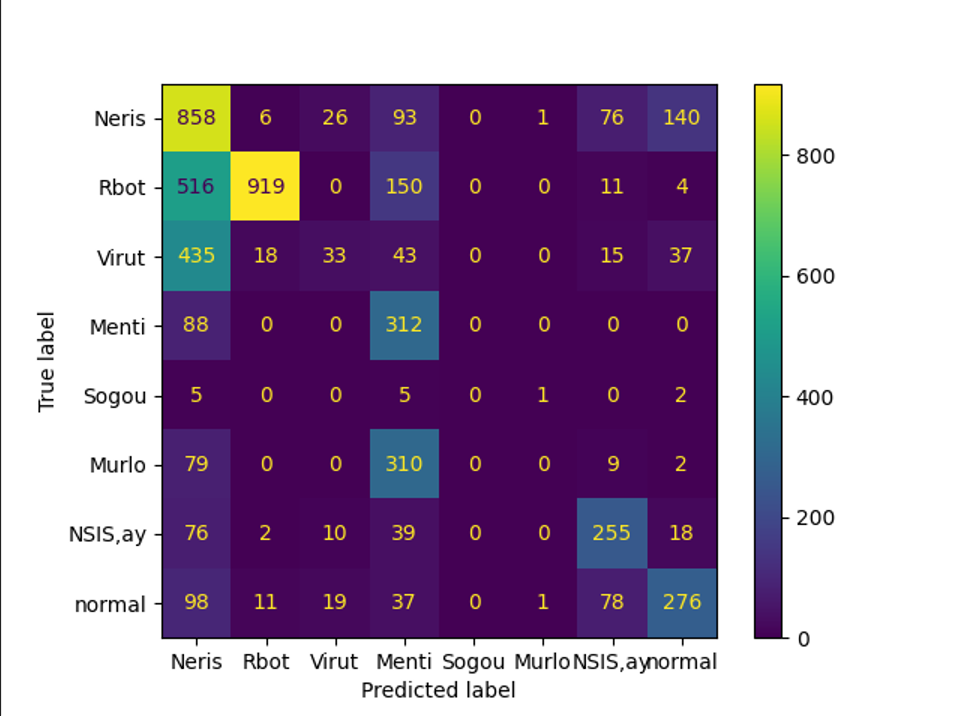


图7

**2）通过logistic回归系数不同僵尸网络攻击特点**

根据图7的混淆矩阵我们选取一些logistic模型能够做出比较准确预测的僵尸网络，并分析这些僵尸网络流量的特点。如图8，左图为用logistic回归系数对Menti僵尸网络分析的结果，从结果可以看出Menti僵尸网络通信的总发包数量一定比较少，然而总的字节数非常多，并且Menti不太可能使用ICMP协议。图8右图为用logistic回归系数对NSIS僵尸网络分析的结果，从分析结果能够看出NSIS僵尸网络不太可能使用ICMP协议，但很可能使用UDP协议或RTP协议，源主机的发送字节数不会太大，但每个包中所含字节数较大

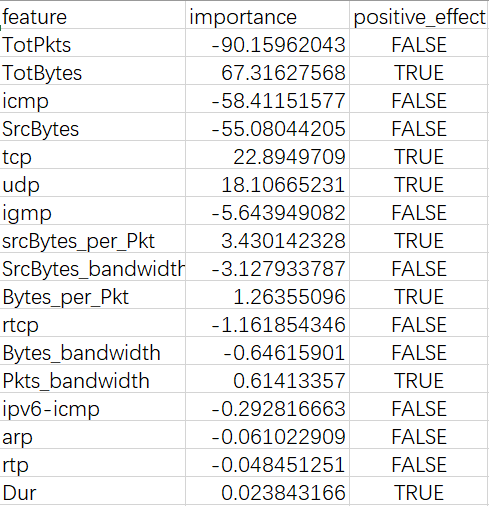
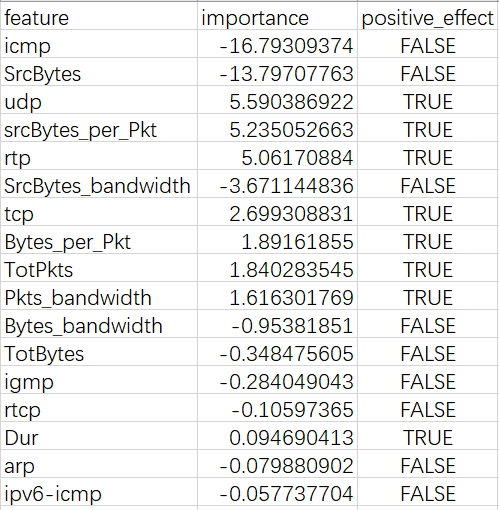
 

图8

**3）决策树混淆矩阵分析**

图9是利用7层决策树对僵尸网络类别进行区分得到的混淆矩阵，可以看出多层的决策树对于僵尸网络的类别能够有很好的区分，许多容易被logistic回归混淆的僵尸网络类别都能被很好辨别出来，这说明我们并不需要[]中所述的用conn.log，ssh.log，x509.log中如此繁多的特征，简单的流量特征就能很好刻画不同僵尸网络的特点。但是多达个叶子节点使得我们无法在此对树进行可视化分析，而且如此大的树会使分析变得异常困难，这也显示了logistic回归的优势，尽管总体表现不一定好，但我们可以通过观察混淆矩阵，分析那些模型预测得不错的类，进行可解释分析，而不是从茫茫个叶子节点中大海捞针。

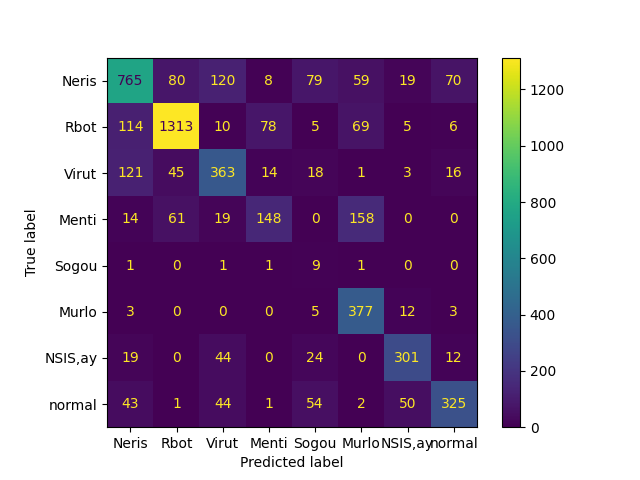
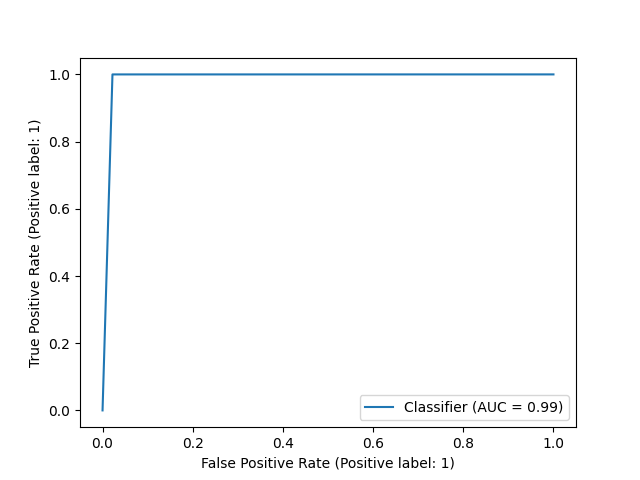
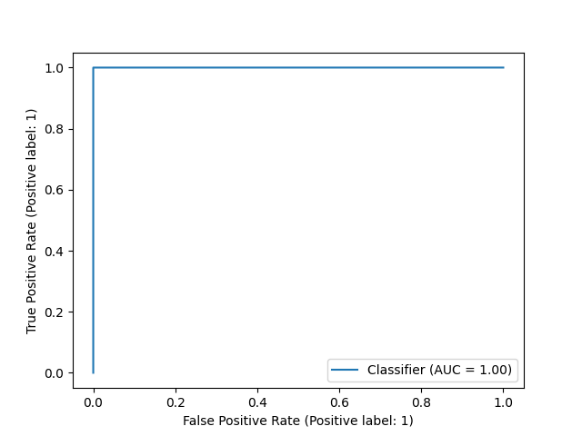


图9



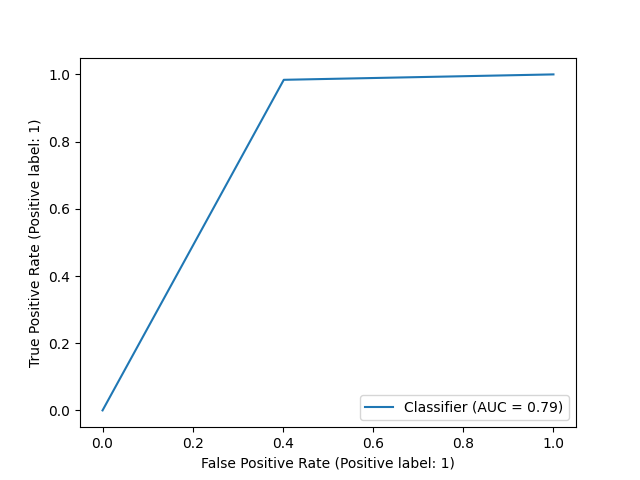
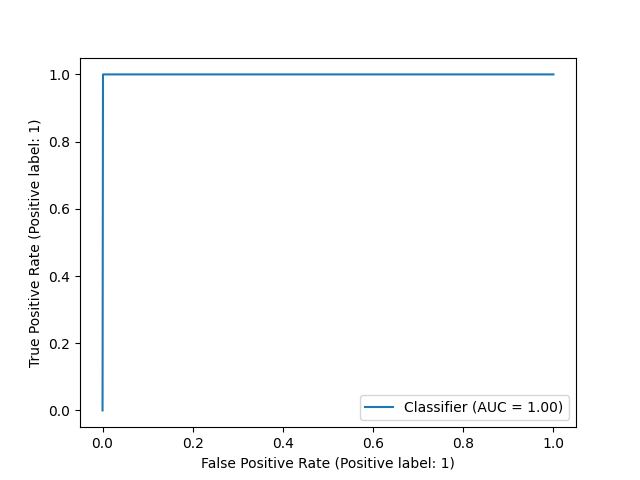


图10

**五、参考文献**

[1]Sharafaldin I, Lashkari A H, Ghorbani A A. Toward generating a new intrusion detection dataset and intrusion traffic characterization[J]. ICISSp, 2018, 1: 108-116.

[2] Garcia S, Grill M, Stiborek J, et al. An empirical comparison of botnet detection methods[J]. computers & security, 2014, 45: 100-123.

[3] Shekhawat A S, Di Troia F, Stamp M. Feature analysis of encrypted malicious traffic[J]. Expert Systems with Applications, 2019, 125: 130-141.

**备注：**所有实验结果图片见实验代码文件夹中的result文件，代码运行见Readme