关于分布式拒绝服务攻击检测的分类算法的评价

摘要——分布式拒绝服务（DDoS）攻击旨在利用恶意流量耗尽目标网络，这对服务的可用性构成了威胁。在过去的二十年里，随着互联网的发展，许多的检测系统被提出，特别是入侵检测系统（IDS），尽管用户和组织在处理 DDoS 时发现该系统很具有挑战性且一直被挫败。虽然，IDS 是保护关键网络不受不断变化的入侵活动问题的第一个防御点，但是它应该一直是最新的以检测任何异常行为，以便保持服务的完整性、机密性和可用性。但是，新的检测方法、技术、算法的准确性很大程度上依赖于是否存在精心设计的数据集，以便通过创建分类器模型进行训练和评估。在这次的工作中，实施了使用主要的监督分类算法的实验，以便从合法流中对 DDoS 攻击进行精确的分类。在所有的分类器中，基于树和基于距离的分类器表现最好。

关键词 - 机器学习，DDoS，逻辑回归，朴素贝叶斯，支持向量机，决策树，随机森林，k-近邻算法。

1.介绍

DDoS 攻击已成为最严重的的网络入侵行为之一，并对计算机网络的基础设施和各种各样的基于网络的服务构成了严重的威胁[1]。它们非常杰出，因为它可以轻易地启动并给组织造成灾难性的损失，而且很难追踪和找出真正的攻击者。DDoS 攻击通过耗尽网络的资源来攻击网络的可用性，从而导致网络拒绝服务，自过去几年以来，网络在数量和容量上迅速增长。攻击持续时间更短、数据量更大的趋势越来越流行[6]。绝大多数现有的工作都使用了如 KDD Cup ’99 或 DARPA 的数据集以检测 DDoSG 攻击。然而，随着时间的推移，网络犯罪和攻击已经以一种巧妙的方式入侵了目标环境。因此，使用具有各种新的攻击特征的最近的数据集来训练分类器，将提高分类器的性能。我们正在利用 CICDDoD2019 数据集进行分析[4]。

我们的工作目标是通过使用 CICDDoS2019 数据集训练模型，实现多个监督分类器以检测 DDoS 攻击。我们的重点是以更高的准确性降低假报，最终有助于提高生产系统的正常运行时间，以及组织的声誉。

2.背景及相关工作

基于 web 服务器日志捕获的特性，例如平均数据包大小、传入比特率与传出比特率、带端口的源 IP 与目标 IP 等，可以检测到网络流量是否异常。主要有两种拒绝服务攻击类型。一种是网络级的 DoS 攻击，它耗尽网络资源且禁止了实际用户的连接，而另一种攻击是应用级的 DoS 攻击，其中服务器资源耗尽，且合法用户请求别拒绝。在 DDoS 攻击中，攻击者可以控制多台叫做僵尸的机器，攻击者可以从中运行一个叫做机器人代码的脚本，并攻击受害者服务器。

主要有两种类别。第一种是反射攻击，另一种是剥削攻击。在反射攻击中，攻击者的身份依然没有被透露，而在剥削攻击中，情况却并非如此。反射攻击和剥削攻击都可以通过应用程序协议、传输层协议或两者的结合来实现。基于 TCP 的反射攻击包括 MSSQL、SSDP，而基于 UDP 的反射攻击包括 CharGen、NTP、TFTP。

[7]中的 Kurniabudi 分析了巨大网络流量的相关和显著特征。Ring 等人已经确定了 15 个不同的属性，以评估单个数据集的实用性[8]。Idhammad 提出了一种基于网络熵估计、聚类、信息增益化和树算法的半监督 ML DDoS 检测算法[9]。[10]中的科研人员提出了 INDB（使用朴素贝叶斯的入侵检测）机制来检测入侵数据包。使用朴素贝叶斯算法的原因是它的可预测性特征。Alenezi 和 Reed 在[11]中对各种疾病进行了广泛的分类。同时对 DoS/DDoS 攻击的困难和特征进行了讨论，并使用了三种不同的分类方法对数据进行了分析。Alpna 和 Malhotra 在 k-近邻和随机森林的帮助下开发了用来检测 DDOS 攻击的架构[12]。Singh 等人开发了一种改进的支持向量机算法，用以检测网络攻击[13]。还有许多设计 DDoS 攻击检测的相关工作。

然而，这些研究大多使用特定的分类算法来评估数据集，并试图使用较旧的数据集（如 KDDCup’99 [2]或 DARPA[3]）专注于优化以获得更好的性能[14 - 16]。在本文中，我们使用最近的数据集 CICDDoS2019[5]对六种不同的分类算法进行了比较分析。

3. 数据集和方法

该数据集有七个超过 10GB 大小数据的 csv 文件。我们应用特征提取算法来寻找最重要的特征，同时执行数据预处理技术，例如数据清理、标准化、去除无穷大值。一旦模型准备就绪，就可以使用测试集通过测量准确度、精确度、召回率、F1 得分、真阳性和真阴性来评估它。如果准确率不能接受，则要对每种分类算法进行优化。此外，还对列车试验的溢出率进行了分析。

DDOS 攻击通常通过僵尸网络或多个机器人产生。因此，虽然在目标服务器上接收数据包时，有多个 IP 地址和 MAC 地址，但是包的长度、流量持续时间、正向包总量等属性使我们识别出它是真正的请求还是恶意请求。为了比较数据包，可以应用数据挖掘技术来测量分类数据分组的概率或发生的情况。在这里，我们使用以下六种机器学习算法对异常流量进行分类：逻辑回归、支持向量机、朴素贝叶斯、k-近邻、决策树和随机森林。

在我们的实验中，我们使用了由新布伦瑞克大学创建的 88 个特征的数据集。该数据集在加拿大网络安全研究所的网站[5]上是开放的。已经收集了不同类型的攻击数据，如 P 我让他马屁、LDAP、MSSQL、UDP、UDPLag 等。如果请求来自合法用户，则它被标记为“良性（Benign）”，否则将被标记为特定的攻击名称。数据集已为分析目的明确创建，而且每天都有被组织起来。每天，CIC 都会记录原始数据，包括来自每台服务器机器的网络流量和事务日志。实际数据集有超过 88 个特征，但是 CIC 本身已经进行了降维，因为他们使用了 CICFlowmeter-V3[17]，并且生成了最重要的 88 个特征以供分析，并提供了 csv 文件。如果有人想通过自己的方式提取功能，他们也已经共享了 PCAP 文件。

我们已经做在数据集上做了两种实验。起初我们从数据集中采样，从每个 csv 文件中随机选择 30000 行，总计 200000 行，作为我们的数据分析样本，这就是我们的不平衡数据集，对于第二个实验，我们从每个 csv 文件数据集中获取相同数量的良性与攻击数据元组，作为一个完全平衡的训练和测试数据集。

表 1 展示了每个文件包含的记录综述与良性标签类的记录总数。更多关于数据集的详细信息可以在[18]中找到。在训练模型之前，数据集中的 IP 地址将被转换为数字整数。

CSV 文件名	总行数	良性行
LDAP	2113234	5124
MSSQL	5775786	2794
NetBIOS	3455899	1321
SYN	4320541	35790
UDP	3782206	3134
乌德普拉格	725165	4068
PortMap	191694	4734
共计	**20364525**	**56965**

表 1 ……

我们选择了单变量选择技术。这是一种统计测试，可用来挑选那些与输出标签关系最强的特征。Scikit 学习库提供了 SelectBest 类帮助我们实现算法，并给出与类标签最相关的特征的结果。我们使用了前 25 个特征来训练我们的模型。为了获得数据集的每个特征的重要性，我们使用了基于树的分类器附带的特征重要性内置类。图 1 说明了前 15 个最重要的特征。

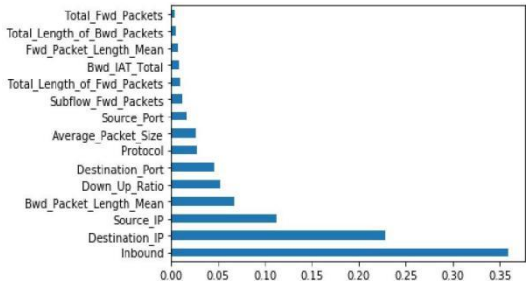

图 1 ……

4. 实验结论及讨论

A. 评估指标

为了评估分类器的性能，我们使用了基于混淆矩阵的主要性能指标。该矩阵包含了关于 ML 模型执行的真实分类和预测分类的信息。为了公平起见，我们还在结果表格中包含了 TP、TN、FP 和 FN 值。如前一节所述，我们在不平衡数据集和平衡数据集上实现了六种不同的机器学习分类算法。我们使用 scikit 学习库运用 python 实现了这两种技术。

B. 实验

我们对每个单独的 7 个 csv 数据文件进行了随机抽样，从每个文件中选择 30K、40K 和 50K 元组，就是为了测量良性流量与攻击流量的比率。实际数据集具有较少数量的良性流量，而在采样时，其本身是有偏差的。当使用不平衡数据训练模型时，与攻击标签相比，平均有 0.5%到 0.7%的良性流量。表 2 展示了类别分布。

样品	攻击（1）	良性（0）
30K	208710	1263
40K	278302	1698
50K	347780	2220

表 2 ……

为了避免对分类模型准确性的偏差问题，我

不平衡数据集	TP	TN	FN	FP	精确度	岁差	召回	F1 积分
决策树	62599	398	3	0	99.99523	1	1	1
朴素贝叶斯	61199	370	31	1400	97.72857	0.6	0.95	0.66
逻辑回归	62619	213	164	4	99.73333	0.99	0.78	0.86
支持向量机	62663	0	337	0	99.46507	0.5	0.5	0.5
K 近邻	62598	401	0	1	99.99841	1	1	1
随机森林	62602	397	0	1	99.99841	1	1	1

表 3 ……

						宏平均量		
平衡数据集	TP	TN	FN	FP	精确度	岁差	召回	F1 积分
决策树	31577	31449	0	0	100	1	1	1
朴素贝叶斯	31387	29278	2290	71	96.25392	0.96	0.96	0.96
逻辑回归	31276	8730	12819	201	79.34185	0.85	0.79	0.78
支持向量机	31577	0	31449	0	50.10154	0.25	0.5	0.33
K 近邻	31477	31549	0	0	100	1	1	1
随机森林	31477	31549	0	0	100	1	1	1

表 4 ……

们从每个 7-csv 文件中选择所有的良性流量，并从攻击流量中随机抽取相同数量的元组创建了平衡数据集。最终，我们从所有文件中收集了 105042 行数据，这些文件具有相同数量的攻击和良性数据。由于这个数字非常小，我们在现有的数据框架中再次追加相同的数据，以将训练集的大小增加到 200k 行以上，这可以与不平衡的数据集相比较。

C. 结果

每个分类器都已使用准确性分数和其他评估指标(如精确度、召回率和 f1 分数)进行了评估。对于不平衡的数据集，每个分类算法的总体精确度展示在表 3 中，对于平衡的数据集，表 4 展示了输出结果。数据是根据五轮观测中的最佳值选择的。

由于不平衡数据集倾向于攻击类别，所以所有分类算法的准确度都很高。但是这并不能帮助我们决定选择出性能最好的 DDoS 攻击检测算法在这里，除了朴素贝叶斯算法之外，所有的算法对不平衡数据集都有很好的处理能力。相反，我们注意到平衡数据集的准确性几乎没有变化。如表 4 所示，基于树的算法，如决策树、随机森林和基于距离的分类算法 k-近邻表现最好，而朴素贝叶斯具有很好的准确率，但其余的分类算法-支持向量机和逻辑回归表现较差。图 2 展示了每种分类算法在不平衡数据集和平衡数据集之间的准确度比较。此外，图 3、4 和 5 分别给出了不平衡数据集和平衡数据集在精确度、召回率和 F-1 得分方面的比较。

5. 今后的工作建议

虽然我们的初步实验结果是有前景的，但这项工作还可以在多个方向上扩展：a）在我们的实验中，由于硬件的限制，我们只使用了 20 多万行。将来我们可以计划选择超过 100 万行的数据集。这将为我们提供更准确的预测训练模型。b）我们可以根据每一种不同类型的 DDoS 攻击进行数据挖掘，因为用 k-近邻可以很好地检测 Portmap，但对于 UDPlag 来说，朴素贝叶斯可能更好，如果证明了这一点，那么我们可以将所有独立的模型合并到一个模型中，以使所有类型的 DDoS 攻击的准确率接近 100%。c）我们可以尝试不同的特征选择技术。

6. 结论

在本文中，我们使用了 CICDDoS2019 数据集，这是一个相当新的数据集，包含了最新的 DDoS 攻击特征。实验使用了主要的监督分类算法进行，以从合法流中准确地分类攻击。在所有的分类器中，决策树、随机森林和 k-近邻算法的分类效果最好。虽然初步的结果是有前景的，但我们计划通过扩展数据集和针对不同类型的 DDoS 攻击来扩展这项工作。我们将把今后的工作重点放在这些方向上。

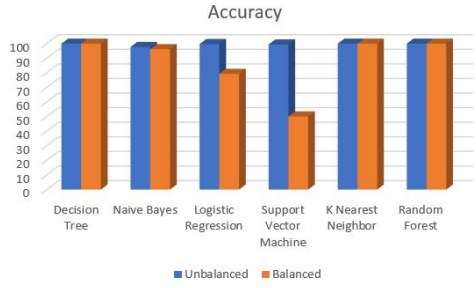

图 2 ……

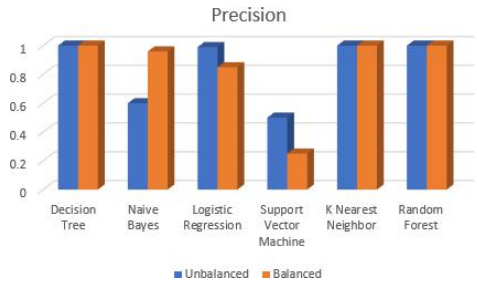

图 3 ……

图 4 ……

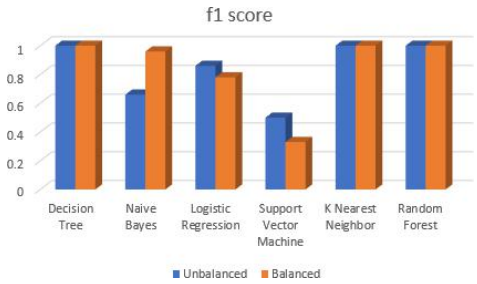

图 5 ……

参考文献

[1]……

实验结果

支持向量机运行结果：

```
准确率为:    0.9966352800020867
TP为:    37270
TN为:    940
FP为:    1
FN为:    128
```

决策树运行结果：

```
准确率为:    0.999947833798482
TP为:    37397
TN为:    940
FP为:    1
FN为:    1
```

逻辑回归运行结果：

```
准确率为:    0.9953311249641358
TP为:    37271
TN为:    889
FP为:    52
FN为:    127
```

随机森林运行结果：

```
准确率为:    0.9917055739586322
TP为:    37080
TN为:    941
FP为:    0
FN为:    318
```

我只下载了数据集中的 CSV-03-11 部分，同时只针对了数据量最小的 Portmap.csv 文件进行训练与测试，不知道是否符合要求，所以复现了 4 个模型。实验中使用的特征字段就是论文中提到的最重要的 15 个，字段和含义如下：

Total Fwd Packets：正向数据包总数（Total packets in the forward direction）

Total Length of Bwd Packets：反向数据包总大小（Total size of packet in backward direction）

Fwd Packet Length Mean：正向数据包平均大小（Mean size of packet in forward direction）

Bwd IAT Total：正向发送两个数据包间的时间（Total time between two packets sent in the forward direction）

Total Length of Fwd Packets：

Subflow Fwd Packets：正向数据包总大小（Total size of packet in forward direction）

Source Port：源端口

Average Packet Size：平均数据包大小

Protocol：协议

Destination Port：目标端口

Down/Up Ratio：下载上传比（Download and upload ratio）

Bwd Packet Length Mean：反向数据包平均大小（Mean size of packet in backward direction）
Source IP：源 IP 地址
Destination IP：目标 IP 地址
Inbound：服务器流入交换机的流量