分布式拒绝服务攻击检测分类算法的评估

摘要：分布式拒绝服务（DDoS）攻击旨在使用恶意流量耗尽目标网络，这对服务的可用性来说是个威胁。随着互联网的发展，在过去的 20 年中，已经有很多检测系统，尤其是入侵检测系统（IDS）被提出，虽然有很多用户和组织发现它在处理 DDoS 时，不断进行挑战，并且被击败。虽然 IDS 是第一个防御点，以保护关键的网络不受新出现的一些侵入式活动的影响，但是他应该时刻保持是最新的，以检测任何异常的行为，以便于保护服务的完整性、机密性和可用性。然而，一些新的检测方式、技术、算法的准确性很大程度上依赖于一些精心设计的数据集，以便于使用创建的分类模型进行训练和评估。在本论文中，已经执行了一些使用主要的监视分类算法的实验，来准确的将 DDoS 攻击和一些合法的流区分开来。在所有这些分类器中，基于树的分类器和基于距离的分类器有最好的效果。

关键词：机器学习、DDoS、逻辑回归、朴素贝叶斯、SVM、决策树、随机森林、K-NN。

1、介绍

DDoS 攻击已经成为一种最严重的网络侵入行为，并且已经对计算机网络的基础设施和和很多基于网络的服务产生了严重的威胁。它们非常有名，因为它们可以被简单的发动，并且对一个组织产生灾难性的损失，除此之外，它非常难以追踪，且难以找到真正的攻击者。DDoS 攻击通过耗尽网络的资源来达到攻击网络的可用性的目标，这会使得网络的服务被拒绝。在近几年来，这种攻击，无论在数量上还是数据量上都有迅速的增加。这一种持续时间短，却带有较大数据量的攻击的趋势已经变得越来越流行。大多数现有的工作都使用了像 KDD Cu ‘99 数据集，或者使用了 DARPA 的数据集来检测 DDoS 的攻击。然而，随着时间发展，网络的犯罪和攻击已经能以一种巧妙的方式来侵入目标环境中。因此，使用一些包含丰富的新颖攻击特征的数据来训练分类器，将会提高分类器的实现效果。在本论文中，我们使用了 CICDDoS2019 数据集来做分析。我们的工作目标是通过使用 CICDDoS2019 数据集来训练模型，并且实现多个监督分类器，以检测 DDoS 攻击。我们的工作重点是减少假阳性，提升高的准确率，并且最终能帮助提升生产系统的正常运行时间，以及组织的信誉。

2、背景和相关工作

基于有网络服务日志所捕获的一些特征，例如平均数据报的大小、传入的比特速率和传出比特速率、源 IP 和目的 IP 地址以及它们的端口等，能够检测网络流量是否异常。现在主要存在者两种类型的拒绝服务攻击。第一种是网络级别的 DoS 攻击，它会耗尽网络资源，从而使得实际的用户连接不可用。而另一种类型的网络攻击是应用层面的网络攻击，在这之中服务资源会被耗尽，因而合法的用户请求被拒绝。在 DDoS 攻击中，攻击者会获得多个叫做“僵尸”的机器的控制权，在这些机器上面，攻击者运行一些叫做机器人代码的脚本，并且攻击受害者服务器。

其中有两种主要的类别。第一种是反射攻击，另一种是利用攻击。在反射攻

击中，攻击者的身份是未知的，而在利用攻击中，攻击者的身份是已知的。而这两种攻击都能被在应用层、传输层或者在两种的结合中实现。基于 TCP 的反射攻击包括了 MSSQL、SSDP，而基于 UDP 的反射攻击包括 CharGen、NTP、TFTP 等。

[7] 中的 Kurniabudi 已经分析了大型网络流量的相关的、重要的特征。Ring 等人已经确定了 15 个不同的属性来范文单个数据集的适用性。Idhammad 描述了基于网络熵估计、聚类、信息增益比以及树算法的实现 DDoS 检测的半监督 ML 方法[9]。[10]中的研究者们提出了 INDB（使用朴素贝叶斯的侵入检测）的机制来检测侵入式的数据包。使用朴素贝叶斯算法的主要原因是它有可预测性的特征。在 [11] Alenezi 和 Reed 提出了 IDS 的一个广泛的的分类。它讨论了 DoS 和 DDoS 攻击的困难和特征，并使用了三种不同的分类方法来分析数据。 Alpna 和 Malhotra 借助了 KNN 和随机森林以开发一个检测 DDoS 攻击的架构。Singh 等人开发了一种开进的 SVM 算法，它被用于检测网络攻击[13]。其中也有很多涉及 DDoS 攻击检测的相关工作。然而，大部分的这些研究都只使用了一种确定的分类算法，并且使用数据集进行评估，并且专注于优化这一种分类方法，且[14-16]使用了一些比较老的数据集，例如 KDDCup’99 [2] or DARPA [3]。而在本论文中，我们针对六种不同的分类方法，使用较新的数据集 CICDDoS2019 [5]，来做了一些比较性的分析。

3、数据集和计算方法

这个数据集有 7 个 csv 文件，包含超过 10GB 的数据。我们使用特征提取算法找到一些最重要的特征并且使用了一些数据处理技术，比如数据清洗、规范化、去除无穷值等。一旦模型确立，就可以使用测试集，通过测量正确性、精确度、召回率、f1 分数、真正负和真负负。如果某一个分类算法的准确性在一个不可接收的范围，那么他必须被优化。除此之外，我们也分析了训练测试溢出比率。

DDoS 攻击一般会通过一个“僵尸网络”或者多个机器人来实现。因此，目标服务器接受数据包时，会有很多的 IP 地址和 MAC 地址，但是一些如包长度、流持续时间、总正向包的总数等属性，都会引导我们识别出这是否是一个真正的请求。为了比较数据包，我们使用了数据挖掘技术来检测数据包分类的的概率或者情况。在本文中，我们使用了以下六种机器学习的算法：逻辑回归、支持向量机、朴素贝叶斯、K 临近、决策树以及随机森林。

为了实现我们的实验，我们使用了一个由 New Brunswick 大学所创建的、由 88 个数据特征的数据集。这个数据集公布在加拿大网络安全研究所的网站上，用户可以公开使用。这个数据集中涉及了不同类型攻击的数据，包括 Portmap, LDAP, MSSQL, UDP, UDPLag 等。来自合法用户的请求会被标注为“良性”，否则会被标注为特定攻击的名称。这个数据集是为了分析而创建，并且每天都会被组织更新。每天 CIC 都会记录一些原始数据，包括每台服务器的网络流量以及时间日志。真实的数据集中包含超过 88 个的特性，但是 CIC 使用了 CICFlowmeter-V3 [17]做了一些降维操作，并且产生了其中最重要的 88 个特征，以进行分析和产生 csv 文件。他们也分享了他们的 PCAP 文件，以便有人想要自己从中抽取特征。

我们对于数据集做了两种类型的实验。一开始我们对数据集做了简化，从每一个 csv 文件中随机抽取了 30，000 行，最终的总和达到 200，000 行，将这些数据作为我们数据分析的样本，是我们的不平衡数据集。而在第二个实验中，我们从每一个数据集中推导出相同数量的良性和攻击数据元组，形成了一个完全平衡的测试和训练数据集。

表 1 展示了每个文件所包含的记录总数和一些普通类，例如标签为“良好”的类。更多关于这个数据集的信息可以在[18]中找到。在训练模型之前，IP 地址被转化为整数形式。

我们选择了单变量选择技术。这是一个统计测试，被用来选择一些于输出标签有着最强关系的特征。scikit-learn 库提供了 SelectKBest 类，它能够帮助我们实现算法，并且给出于我们的类标签有着最大关联性的特征。我们使用了前 25 个特征来训练我们的模型。为了获得每数据集中每一个特征的重要性，我们使用了基于树的构造的特征重要性内置类。图 1 展示了最重要的 15 个特征。

4、实验结果和分析

A、评估指标

为了评估不同分类器的表现，我们使用了基于混淆矩阵的重要性能指标。这个矩阵包含了一些由 ML 模型执行的、有关于真实和预测分类的信息。公平的来说，我们在我们的表中包含了 TP、TN、FP 和 FN 的值。正如上一节所提到的，我们基于不平衡的数据集和平衡数据集实现了六种不同的机器学习分类算法。我们也用 python 使用 scikit-learn 库来实现这两种方法。

B、实验

我们在 7 个 csv 数据文件的每一个独立文件中，随机的取样，并且每一个文件选取 30K、40K 和 50K 的元组来测量良性流量和攻击流量的比例。实际的数据集会有一个更小数量的良性流量，并且它被采样时，它自己是有偏见的。平均的来说，当我们使用不平衡数据训练模型时，平均的良性流量为 0.5%~0.7%。表格 2 显示了它的分布。

为了避免这样的分类模型正确性的偏差问题，我们创建了平衡的数据集，在其中我们从 7 个 csv 文件中的每一个选取相同的良性流量元组，并从攻击流量中随机抽取相同数量的元组。我们最终从所有文件中选取 105042 行，其中相同的良性和攻击数据。由于这个是一个非常小的数字，我们在现有的数据框架中再一次增加相同的数据，来增加训练集的大小，使得训练集超过 20K 行，可与不平衡数据集比较。

C、结果

每一种分类器都已经使用正确性分数和一些其他的评估指标，如精确度、f1 的分数等。每一个分类算法的总的正确性被显示在不平衡数据集的表格 3 中，而在表格 3 中，展示了平衡数据集的输出结果。数据选择根据无论观察的最小值进的行。

由于不平衡数据集更加偏向攻击类，预测分类的算法的正确性也会比较高。然而这不能帮助我们选择 DDoS 攻击检测的性能最好的算法。我们的实验中，除了朴素的贝叶斯之外，所有的算法对于不平衡数据的处理都非常好。相反的是，我们意识到不同方法在平衡数据集的准确性基本没有什么变化。就像表格 4 中所显示的那样，基于树的算法例如决策树、随机森林、以及基于距离的分类算法 K-NN 表现的最好，而朴素的贝叶斯算法算法能够实现良好的精度，但其他的一

些分类算法，如 SVM 和逻辑回归，则分类结果较差。图片 2 展示了每一种分类算法分别使用平衡数据集合和平衡数据集的得分比较。而图片 3、4、5 展示了不同分类算法的使用平衡数据集和不平衡数据集的各自的精确度、召回和 F-1 分数的比较。在分析了输出之后，我们发现基于树的分类算法，如决策分析与随机森林，以及基于距离的分类算法在两种数据集上都有最好的效果，并且都得到了几乎 100%的正确率。即使考虑了一些其他的度量标准时，这三个分类器也有最好的效果。然而，当每一个分类器的参数发生变化时，可以注意到它们的结果都会产生微小的变化。我们尝试给出了每一种算法的最好表现情况。

5、未来工作建议

我们的实验结果非常鼓舞人心，并且它可以从很多的方向进行拓展。a）在我们的实验中，由于硬件的限制我们只使用了略超于 200，000 行的数据。我们计划在未来选择约有 100 万行的数据集进行实验。这会给我们们的预测带来一个更加精确的模型。b）我们可以基于不同类型的 DDoS 攻击做数据挖掘，因为 K-NN 可能以较高的效率检测到 Portmap，但是对于 UDPPlag 来说，朴素的贝叶斯可能会更好。如果说上述结论可以被证明的话，我们可以将所有的独立的模型进行合并成为一个模型，并且对于所有的 DDoS 攻击都能够得到接近于 100%的精确度。c）我们可以尝试不同的特征选择方式。

6、结论

在本论文中，我们使用了 CICDDoS2019 数据集，它是一个相当新的数据集，并且包含了大部分的 DDoS 的攻击前面。一些使用主要的监督分类算法的实验已经被执行，他们能够将攻击的流量和合法流量区分开来。当所有的分类器的算法进行比较时，决策树、随机森林和 K-NN 表现的最好。虽然初步的结果是很有希望的，我们准备用扩展的数据集并针对不同类型的 DDoS 攻击以扩展我们的工作。我们未来的工作将专注于上述各方面。

参考文献