DOI:10.13718/j.cnki.xsxb.2021.01.005

基于并行积累排序算法和主动学习的 DDoS 攻击检测①

王 慧[1], 张学军[2]

1. 柳州职业技术学院 艺术学院, 广西 柳州 545006; 2. 北京控制与电子技术研究所, 北京 100038

摘要: 为了在高速网络环境下对大容量网络流量进行准确和快速的分类, 以检测分布式拒绝服务(Distributed Denial of Service, DDoS)攻击, 本文提出一种基于并行积累排序算法和主动学习的 DDoS 攻击检测算法. 该技术采用并行积累排序算法对流量特征进行积累排序来选择最佳特征子集, 通过专家模块以无监督的方式选择适当的实例来训练用于检测 DDoS 攻击流量的支持向量机(SVM)二值分类器, 从而实现从数据集中选择小批量训练样本来产生高精度的网络流量分类. 实验结果表明, 与现有方法相比, 本文算法在分类准确率和执行速度方面均优于现有方法.

关 键 词: 并行积累排序; 主动学习; 支持向量机; DDOS 攻击

中图分类号: TP393 **文献标志码:** A **文章编号:** 1000-5471(2021)01-0025-07

分布式拒绝服务(Distributed Denial of Service, DDoS)攻击由于攻击签名不断变化而很难防御, 对各种业务和企业构成了严重威胁[1-2]. 快速有效的网络流量识别和分类可以显著提高网络安全, 由于传输数据的大小不断增加以及可用的应用程序的多样性, 必须通过流量分析进行流量优先级排序和诊断监控[3-5]. 信息多样性或传播对网络流量分类来说是一个很大的挑战, 信息传播意味着每种类型的流量都可以具有独特的特征或统计属性. 集体分类指使用所有可能的信息对一组相互关联的对象进行分类, 为了执行集体分类任务, 需要为流量实例的初始群体检索类别标签, 并在下一轮分类中使用这类标签. 因此, 在对整个业务进行分类之前需要用所需的信息来标记部分被选择的实例, 并且确定它们对于不同类别的归属. 基于初始信息可以成批对网络流量的所有其他剩余实例进行分类[6].

主动学习是半监督机器学习的一种特例[7-8], 其中学习算法能够交互式地查询用户(或某些其他信息源)以获得新数据点上的期望输出, 被称为最佳实验设计[9]. 虽然存在未标记数据丰富的情况, 但是手动标记这些数据成本非常昂贵, 而学习算法可以主动地向用户、教师或专家查询标签, 这种类型的迭代监督学习称为主动学习. 由于学习者选择示例, 因此用于学习概念的示例数量通常会远远低于正常监督学习所需的数量. 本文使用具有较少训练实例的主动学习法来处理大量的网络流量.

能够正确和快速地检测 DDoS 攻击是网络安全需要解决的关键技术. 近年来, 有关 DDoS 攻击检测系统的研究已取得若干成果. 文献[10]提出了一种基于多级自动编码器特征学习的高效 DDoS 攻击检测技术, 该技术以无监督方式学习多层次的浅层和深层自动编码器来对训练和测试数据进行编码, 以用于特征再生, 通过使用有效的多核学习算法组合多级特征来学习最终的统一检测模型. 文献[11]比较了集中式和分布式特征选择方法, 该方法垂直或水平地划分数据集, 可以在显著减少运行时间的情况下获得更高的分类性能. 文献[12]提出了一个快速最小冗余最大相关性算法, 并在几个不同的平台上得到了实现, 即用于

① 收稿日期: 2020-04-17
基金项目: 2020 广西高校中青年教师科研基础能力提升项目(2020KY31023, 2020KY31024).
作者简介: 王 慧, 硕士, 副教授, 主要从事信息安全研究.

顺序执行的中央处理器(CPU)、用于并行计算的图形处理器以及用于使用大数据技术进行分布式计算的Apache Spark.

结合文献[11]划分数据集的方法和文献[12]快速最小冗余最大相关性算法的优点，本文提出一种基于并行积累排序算法和主动学习的DDoS攻击检测技术. 该技术在GPU的核心之间分配海量网络流量数据集的计算负载，并将特征选择方法局部应用于每个核心，可以处理大型数据集，并在不影响质量的情况下近乎实时地对其进行处理. 本文首先在并行计算环境中以积累排序方式来对网络流量进行排序，以此选择最佳特征子集. 为了大量处理网络流量，通过专家模块以无监督的方式选择适当的实例来训练SVM二值分类器，从而实现从大容量网络流量选择小批量训练样本产生高精度网络流量的分类目的. 实验结果显示，本文算法在执行时间和分类准确度性能方面优于其他方法.

1 并行积累排序(PCR)算法

并行积累排序(Parallel Cumulative Rank，PCR)算法在GPU核心之间分配海量网络流量数据集的计算负载，并将特征选择方法局部应用于每个核心，将所选择的特征放在一起并在累积的基础上进行排序，以此实现在不影响质量的情况下近乎实时地处理大型数据集. 本文PCR框架如图1所示，PCR算法有3个主要任务：特征排序、特征选择和数据分类.

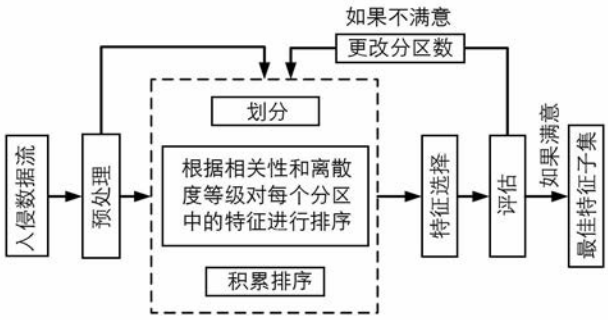

图1 本文设计的PCR框架

1.1 特征排序

特征排序首先对数据集每个分区的特征进行排序，特征的等级定义了特征的相关性和非冗余性，通过组合每个部分的单个排序来计算全局或累积排序，等级越高被包括在用于分类的特征子集中的可能性越高. 特征排序有3个并行执行的子任务：预处理、数据分区和平行排序.

(1) 预处理

选择网络流量的5个属性来构建原型，即每个流量样本的源IP地址(F1)、目的IP地址(F2)、源端口号(F3)、目的端口号(F4)和帧长度(F5). 从包含大约100万个数据包的流量数据中获得这些属性的值.

(2) 数据集分区

为了进行分布式并行计算，本文对网络流量数据进行垂直划分. 本文改变分区的数量，计算每个分区中每个特征的排序，并跟踪顺序环境和并行环境中的执行时间.

(3) 平行排序

每个分区都在具有GPU的机器上处理，没有任何重叠. 对每个分区单独执行秩计算，同时考虑相关性和冗余性来计算属性的等级，使用相关性排序和离散度等级两种方法来推导每个特征的等级，根据样本集合中属性的唯一性找到特征的等级. 相关性排序提供属性之间的水平唯一性，离散排序提供属性之间的垂直唯一性.

定义1：相关等级R_i被定义为特征f_i的等级，由相关系数Cor_i相对于给定类别s'的特征子集C_i的所有其他特征给出. 该系数给出了特征f_i与来自相同子集的所有其他特征的相关性的度量.

定义2：散度等级D_i被定义为特征f_i的等级，由散度系数或散度指数$Disp_i$相对于给定类别s'的特征子集C_i的所有其他特征给出. 此系数提供了特征f_i的实例如何聚集(同构)或分散的度量.

定义3：累积等级Cum_i被定义为特征f_i相对于给定类别s'的特征子集C_i的所有其他特征的全局等级. 该等级通过考虑特征子集中f_i的特征s'的相关等级和散度等级两者的累积等级来给定类别C_i.

相关性排序：本文使用相关性度量来计算一个属性与所有其他属性上的相关性，属性的等级可以基于给定类别的相关系数来计算，皮尔逊相关系数p是两个变量x和y之间线性相关的度量，p从$-1\sim+1$取值，当p值为$+1$时表示两个变量线性相关，当p值为-1时表示两个变量负相关，当p值为0时表示两个变量彼此不相关. 对于m个特征的集合，通过计算每个特征的$m-1$个p值来找到该类的相关性. 一个特征

的平均 p 值给出了该特征相对于所有其他特征的相关性，因此可以建立所有特征之间的等级.

$$p=\frac{(n\sum(x\times y)-(\sum x)\times(\sum y))}{\sqrt{(n(\sum x^2)-(\sum x)^2)\times(n(\sum y^2)-(\sum y)^2)}} \tag{1}$$

离散度等级：离散度等级用于量化属性的一组观察实例相对于给定类别的离散程度，可以提供数据集中属性的值相对于该属性平均值的离散系数，即它是一种用于评估一组观察到的事件是紧密聚集还是分散聚集的度量. 较高的分散值意味着较大的分散. 本文通过将离散度等级应用于对一个属性的排序来获得更多的相关性和非冗余特征. 特征的离散系数 q 由式(2) 给出.

$$q=\frac{\text{方差}}{\text{均值}}=\frac{\sigma^2}{\mu} \tag{2}$$

其中，σ^2 为方差，μ 为均值.

累积排序：累积排序被定义为在考虑所有分区的情况下将相关性和离散度等级组合，以便更好地计算给定类属性的相关性后得到的等级. 相关等级为我们提供了水平相关性，而离散等级为该类提供了属性的垂直相关性. 单个属性两个等级的平均值为数据集单个分区的所有属性提供一个等级，因此累积所有分区的等级以产生全局等级. 随着垂直分区数量的变化，相关等级和离散等级均会变化，累积等级也会改变. 值得注意的是，如果改变分区的数量，固定数据集的一组要素的等级不会保持不变，此时分类准确性也会变化. 对于固定数据集，通过调整垂直分区数量以获得更好的分类准确度，而并行计算有助于轻松地进行这种调整.

1.2　特征选择

特征选择通过从数据集中去除无关和冗余特征来选择相关特征子集以建立学习模型，包括从所有可能的子集中寻找最佳特征子集的搜索过程. 应用评估度量来估计和评估给定类别的每个特征子集的相关性，可以通过包装器、过滤器和嵌入方法 3 种方式选择或排序.

特征排序过程为本文提供每个特征的等级，本文使用排序中的分界点来选择最佳特征的子集，该分界点可以通过经验结果以启发方式选择，本文根据截止阈值选择 3 个最佳属性进行分类.

1.3　数据分类

为了评估本文方法的性能，本文选择了 5 种著名的分类算法来检查特征排序过程的准确性：精细 k 最近邻(KNN)、线性判别、Logistic 回归、Boosting 和复杂决策树. 从每个数据集中取一个包含约 10 万个数据包的分区，应用特征排序过程来获得前 3 个特征，并使用这些特征进行分类.

以下命题对于平行累积排序(PCR) 正确.

命题 1：具有给定类别 C_i 的高相关等级特征 s' 的子集与该类相关，即 $s'\odot C_i$，其中 $\odot$ 表示由于高相关性引起的相关性关系.

说明：相关秩用于评估两个或多个变量之间关系的重要性. 假设 f_1，f_2，…，f_n 是给定类 C_i 的特征 s' 的子集中的特征或属性. 如果对于给定类 C_i 的 f_1 的皮尔逊相关系数接近 $+1$ 或 -1，那么 f_1 与同一类特征子集的所有其他特征相关，具有秩越高的特征相关性越高.

命题 2：具有给定类 C_i 的低离散度等级的特征 s' 子集与该类相关，即 $s'\odot C_i$，其中 $\odot$ 表示低离散度的相关性关系.

说明：离散系数是一种度量，用于评估给定类的属性值或实例是如何聚类或分散的. 假设 f_1 是给定类 C_i 的特征 s' 子集中的特征或属性，离散指数用于测试属性观测值的均一性，当色散系数较低时，类别 C_i 的特征 f_1 的实例称为“色散不足(更均匀)”，否则为“色散过度(不均匀)”. 系数值越低，给定类别 C_i 的色散秩越高，相同类的同构属性更具有相关性.

命题 3：对于给定的类 C_i 具有高累积秩的特征 s' 子集是相关的.

说明：假设 f_1 是给定类 C_i 的特征 s' 子集中的特征或属性，如果它具有更高的相关等级和更高的离散度秩，则类 C_i 的累积秩或相关性比来自相同特征子集的所有其他属性更高.

2　网络流量分类中的主动学习

本文设计了一种主动学习方法来对网络流量进行分类以检测 DDoS 攻击，使用适当的训练样本，由专

家模块在批处理模式下从未标记的数据池中进行识别. 图2给出了本文设计的主动学习框架，整个过程可以根据不同的组件进行可视化.

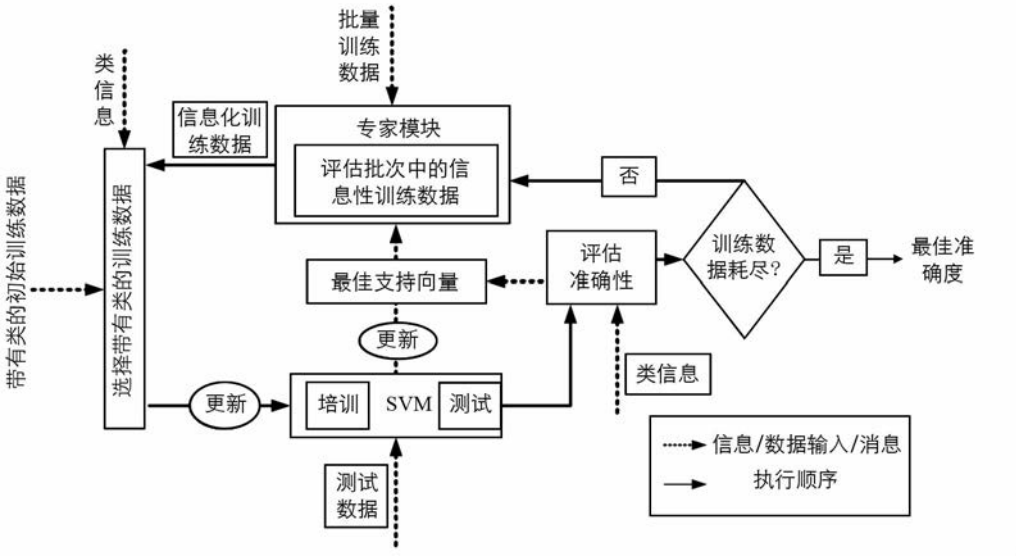

图2 主动学习框架

设A是所考虑的全部数据集，DDoS攻击时DDoS流量嵌入到整个流量中. A将包括已知具有特定恶意活动的所有网络流量实例，以及可能想要测试该活动的所有其他实例. 在每次迭代 i 期间，A被分成3个子集：标签已知的数据点 $A_{k,i}$、标签未知的数据点 $A_{u,i}$ 和需要从训练数据中学习来标记的 $A_{u,i}$ 的 A 子集.

2.1 数据组件

本文使用4个数据集：MIT-DARPA、CAIDA-2007、ISCX和TUDDoS，使用Editcap和Tshark对每个数据集包含大约100万个数据包进行预处理以获得这些属性的值. 选择网络流量的5个属性：源IP地址、目的IP地址、源端口号、目的端口号和每个流量样本的帧长度，源IP地址和目的IP地址已转换为十进制值.

训练数据：训练数据由正常流量实例和DDoS流量实例组成，使用5个属性进行描述，每个实例的标签来自标签池. 这些数据用于训练SVM二值分类器.

测试数据：测试数据也包括正常流量实例和DDoS流量实例，每个实例的标签都在标签池中. 训练好的SVM分类器对测试数据进行分类.

标签池：本文将培训和测试数据的所有标签保存在单独的池中，并且对每个培训和测试数据的实例都有一个准确的标签映射，没有任何错误. 当需要提供适当的实例以更新SVM分类器的训练数据时，将用于训练数据的标签提供给专家模块.

最佳可能支持向量：该数据组件是一组支持向量，它们被迭代更新. 在训练数据完全耗尽之后，该组件具有用于对测试数据进行分类的最佳支持向量集.

2.2 流程组件

对于主动学习，本文根据框架开发了5个核心处理组件：训练数据选择、专家模块、专家模块策略、SVM分类器和评估准确性. 首先随机选择几个训练样本，并用训练样本训练SVM，在测试数据上测试SVM，使用公式(3)来查找分类精度需要测试数据上的标签，如果准确性大于最佳准确性，则更新最佳准确度，并使用支持向量集更新最佳支持向量集. 如果训练数据已用尽则停止寻找准确度，否则将最佳向量集输入专家模块，调用专家模块获取下一批信息丰富的训练样本.

训练数据的选择：为了启动主动学习过程，本文随机从训练数据中选择一些流量实例进行学习，采用SVM来计算 n 支持向量. 专家模块根据来自该组件的请求，使用这些向量从标签池中提供具有类别标签的适当训练样本，在每次迭代中专家系统评估来自训练数据的 $3\times n$ 实例. 在SVM训练阶段，训练数据选择组件更新适当的实例.

专家模块：一批训练数据中的样本数量是分类器前一轮 SVM 训练产生的支持向量总数的 3 倍，专家模块根据专家模式策略处理该批训练数据以找到所需的样本. 选择的数据样本和类与前一组训练数据合并以便在下一轮进行进一步训练，从而得到更好的边界线和更新的支持向量集.

专家模块策略：设 n 是在 SVM 分类器第$(i-1)$次训练迭代中生成的支持向量数量，从主训练数据集中汇集一批大小为 $3\times n$ 的实例，以便在专家模块中进一步评估以找到适当的实例，从 SVM 分类器提供的每个支持向量中计算批次中每个样本的欧几里德距离. 选择距每个支持向量最短和最长距离的样本与支持向量机分类器的训练数据合并，以创建第 i 次迭代的训练数据. 因此，策略性选择的样本数量小于或等于 $2\times n$. 对于这些选定的样本，本文使用存储在标签池中的原始类/标签. 欧几里德距离可以用其他距离度量代替.

SVM 分类器：SVM 分类器由培训和测试两个阶段组成. 选择或更新的训练数据用于学习，并且在该训练阶段用于分类的超平面构造. 将生成的支持向量提供给专家模块以便适当选择下一次迭代学习所需的实例. SVM 分类器使用学习阶段生成的超平面对测试数据进行分类.

评估准确性：使用分类标签与实际类标签之间的差异来计算所设计模型的准确性. 当 SVM 分类器将测试数据分类为两个单独的类时，使用公式(3)来查找分类精度需要测试数据上的标签.

$$TRUE\_\text{accuracy}_i = 100 - \left(\frac{E_l}{T_l}\times 100\right) \tag{3}$$

其中，E_l 为错误标签，T_l 为总的测试标签.

3　实验结果与分析

为了对实验进行评估，本文在配置为 Intel(R)Core(TM)i5-3320CPU@2.30 GHz 处理器、64 GB RAM 的 64 位 Windows 10 操作系统上，使用 MATLAB R2016a 的并行计算工具箱，选择分布式模型单个机器中存在的工作人员来进行实验，假设从 CPU 向 GPU 单个工作节点发送业务数据所需的通信时间是恒定的. 使用 4 个数据集 MIT-DARPA、CAIDA-2007、ISCX 和 TU-DDoS 中的数据包对 3 组数据包进行预处理，在每组数据包中，本文合并 60%的正常流量和 40%的 DDoS 攻击流量. 通过逐一改变初始训练数据、分批训练数据、训练数据总量和测试数据总量 4 列的条件来找出对分类精度的影响.

图 3 给出了所有数据集下改变初始训练数据实例数量时本文算法的准确度，从而找到初始实例量以获得高精度. 提供高精度的少量初始实例使主动学习过程更有效率. 从图 3 可以看出 4 个数据集 MIT-DARPA，CAIDA-2007，ISCX 和 TU-DDoS 的初始实例分别为 70，60，50 和 50.

图 4 在保持各数据集初始实例数量不变的情况下，改变下一次迭代的批量处理实例数，从而找出合适的批量处理实例数量. 如果批次中较低数量的实例可以提供比批次中较高数量的实例更高的准确性，则专家模块使用支持向量查找适当的实例进行进一步分类的工作量会降到最低. 从图 4 可以看出 4 个数据集 MIT-DARPA，CAIDA-2007，ISCX 和 TU-DDoS 的批量处理实例数分别为 100，300，300 和 700.

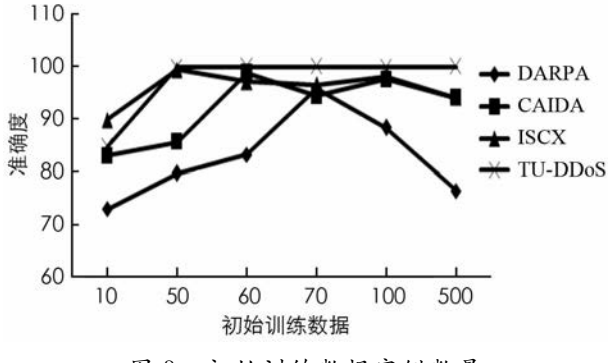

图 3　初始训练数据实例数量不同时各数据集的准确度

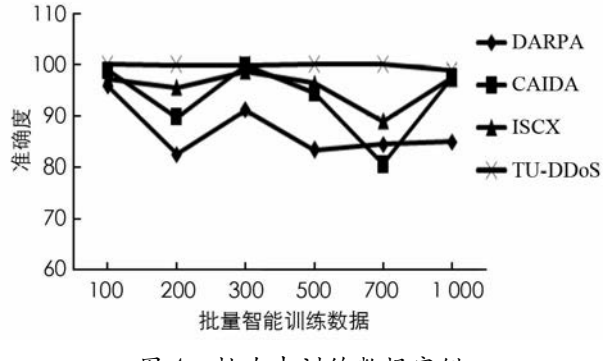

图 4　批次中训练数据实例数量不同时各数据集的准确度

然而，批处理中实例数量较多时通过整个训练数据集运行的迭代较少. 因此，本文在图 5 中改变了训

练实例的总数，在批次中的实例数与训练数据集中的实例总数之间进行完美的调整，可以减少遍历整个训练数据的迭代次数. 从图5可以看出在固定前两列不变的情况下，4个数据集MIT-DARPA,CAIDA-2007,ISCX和TU-DDoS数据实例总数分别为2000,10000,5000和2000.

图6在将前3列固定以获得更高精确度特定设置的情况下更改了测试数据中的实例数，以此来研究测试数据中较高数量的实例的准确性如何变化. 从图6可以看出对前3列进行固定后，精确度不会随着测试数据中实例数的增加而有太大变化，这从主动学习的角度来看是需要的.

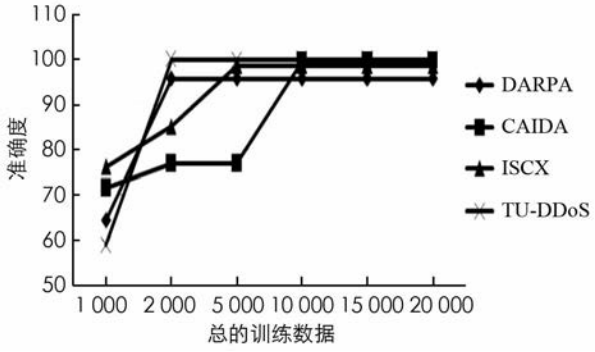

图5 训练数据实例总数不同时数据集的准确度

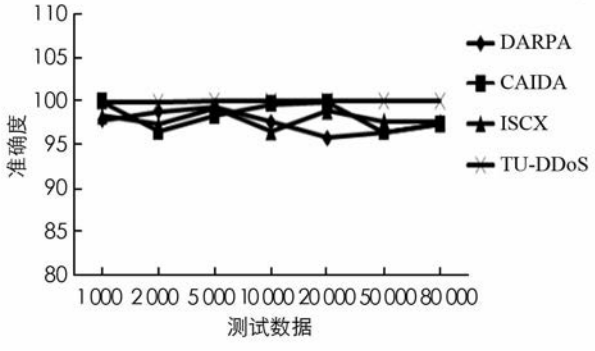

图6 测试数据实例总数不同时各数据集的准确度

表1给出了本文算法同基于模糊性的半监督学习方法和使用支持向量机(SVM)训练模型的性能比较. 从表1可以看出本文算法有比较高的准确度和较快的执行速度，这是因为本文使用并行积累排序算法来对数据集属性进行排序以找到最优特征子集，使用并行计算和具有较少训练样本的主动学习方法来处理大量的网络流量.

表1 各算法性能比较

方 法	样本数	准确度/%	运行时间/s
半监督学习法[13]	10 000	84.1	287.1
SVM训练模型[14]	10 000	98.6	318.4
本文算法	10 000	99.96	15.3

4 结 语

为了对大量的网络流量进行正确和快速地分类以检测DDoS攻击，本文采用基于并行积累排序算法和主动学习的DDoS攻击检测方法. 该方法通过并行积累排序算法对数据集的属性进行排序来寻找最佳可能的特征，使用并行计算方法来处理大量的网络流量，并讨论了主动学习的重要性，通过专家模块以无监督的方式选择适当的实例来训练用于检测DDoS攻击流量的SVM二值分类器，以此实现从数据集中选择小批量训练样本来产生高精度的网络流量分类. 实验结果表明，本文算法在处理大流量数据分类时，在训练样本较少的情况下提供了更好的分类准确率和更快的速度. 未来的工作是通过结合软计算和其他技术开发一种由主动学习支持的模糊推理来扩展PCR，以便能够对大量数据空间的特征进行排序以建立其通用性.

参考文献:

[1] 汪 洋，伍忠东，朱 婧. 基于深度序列加权核极限学习的入侵检测算法[J]. 计算机应用研究，2020，37(3)：829-832.

[2] SINGH P K，JHA S K，NANDI S K，et al. ML-Based Approach to Detect DDoS Attack in V2I Communication Under SDN Architecture[C]//TENCON 2018-2018 IEEE Region 10 Conference. Jeju：IEEE，2018.

[3] PACHECO F，EXPOSITO E，GINESTE M，et al. Towards the Deployment of Machine Learning Solutions in Network Traffic Classification：a Systematic Survey[J]. IEEE Communications Surveys & Tutorials，2019，21(2)：1988-2014.

[4] 燕昊昊，韩国栋，黄雅静，等. 非平衡网络流量识别方法[J]. 计算机应用，2018，38(1)：20-25.

[5] WANG P，YE F，CHEN X J，et al. Datanet：Deep Learning Based Encrypted Network Traffic Classification in SDN Home Gateway[J]. IEEE Access，2018，6：55380-55391.

[6] 刘　敏，滕　华，何先波. 基于核函数的软件定义网络 DDo S 实时安全系统 [J]. 计算机应用研究，2020，37(3)：843-846，850.

[7] BARTHOLOMEW J B，JOWERS E M，ROBERTS G，et al. Active Learning Increases Children's Physical Activity across Demographic Subgroups [J]. Translational Journal of the American College of Sports Medicine，2018，3(1)：1-9.

[8] SHEKHAR P，PRINCE M，FINELLI C，et al. Integrating Quantitative and Qualitative Research Methods to Examine Student Resistance to Active Learning [J]. European Journal of Engineering Education，2019，44(1/2)：6-18.

[9] MELNIKOV A A，POULSEN NAUTRUP H，KRENN M，et al. Active Learning Machine Learns to Create New Quantum Experiments [J]. PNAS，2018，115(6)：1221-1226.

[10] YAN B H，HAN G D. Effective Feature Extraction via Stacked Sparse Autoencoder to Improve Intrusion Detection System [J]. IEEE Access，2018，6：41238-41248.

[11] MORÁN-FERNÁNDEZ L，BOLÓN-CANEDO V，ALONSO-BETANZOS A. Centralized Vs. Distributed Feature Selection Methods Based on Data Complexity Measures [J]. Knowledge-Based Systems，2017，117：27-45.

[12] RAMÍREZ-GALLEGO S，LASTRA I，MARTÍNEZ-REGO D，et al. Fast-mRMR：Fast Minimum Redundancy Maximum Relevance Algorithm for High-Dimensional Big Data [J]. International Journal of Intelligent Systems，2017，32(2)：134-152.

[13] ASHFAQ R A R，WANG X Z，HUANG J Z，et al. Fuzziness Based Semi-supervised Learning Approach for Intrusion Detection System [J]. Information Sciences，2017，378：484-497.

[14] CAO J，FANG Z，QU G，et al. An Accurate Traffic Classification Model Based on Support Vector Machines [J]. International Journal of Network Management，2017，27(1)：1-15.

DDoS Attack Detection Based on Parallel Accumulation Ranker Algorithm and Active Learning

WANG　Hui[1]，　ZHANG Xue-jun[2]

1. School of Art，Liuzhou Vocational and Technical College，Liuzhou Guangxi 545006，China；

2. Beijing Institute of Control and Electronic Technology，Beijing 100038，China

Abstract：To classify accurately and quickly large capacity network traffic in high-speed network environment to detect distributed denial of service (DDoS) attacks，a DDoS attack detection algorithm based on parallel cumulative ranker algorithm and active learning has been proposed in this paper. In this technique，the parallel accumulation ranker algorithm has been used to accumulate and rank the traffic features to select the best feature subset，and the expert module selects the appropriate examples in an unsupervised way to train the support vector machine binary classifier for detecting DDoS attack traffic，so as to select a small number of training samples from the data set to generate high precision network traffic classification. Experiments show that compared with the existing methods，the proposed algorithm is superior to the existing method performance in classification accuracy and execution speed.

Key words：parallel accumulation rank；active learning；support vector machine；DDoS attack

责任编辑　夏　娟