一种基于数据挖掘的拒绝服务攻击检测技术

高 能 冯登国 向 继

(中国科学院研究生院信息安全国家重点实验室 北京 100049)

摘 要 提出了一种新的、基于数据挖掘的 DoS 攻击检测技术——DMDoSD,它首先利用 Apriori 关联算法从原始网络数据中提取流量特征,然后利用 K-means 聚类算法自适应地产生检测模型,这两种算法的结合能够实时地、自动地、有效地检测 DoS 攻击. DMDoSD 除了向现有的 IDS 发出攻击报警外,还进一步利用关联算法分析异常网络数据包,确定攻击特征,为 DoS 攻击的防御提供支持.

关键词 拒绝服务攻击;聚类算法;关联算法;实时检测

中图法分类号 TP393

A Data-Mining Based DoS Detection Technique

GAO Neng FENG Deng-Guo XIANG Ji

(*State Key Laboratory of Information Security, Graduate University of Chinese Academy of Sciences, Beijing* 100049)

Abstract Denial of Service (DoS) is a type of frequent network attack which can severely impact the availability of networks and services. DoS usually utilizes packet-attribute spoof techniques to confuse present IDSs such as snort. Typically, the spoof techniques minimize effective and automatic DoS attacks detection. A novel technique based on data mining to detect DoS attacks in real-time called DMDoSD is presented. First, the Apriori association algorithm extracts traffic patterns from empirical network data and subsequently the K-means cluster algorithm adaptively generates a detection model. By combining these two algorithms, DoS attacks can be detected swiftly, automatically and effectively as they arise. In addition to the alerts typically sent out by IDSs, DMDoSD also determines signatures of malicious packets automatically to help to react to DoS attacks.

Keywords DoS (Denial of Service) attack; cluster algorithm; association algorithm; real-time detection

1 引 言

拒绝服务(Denial of Service,DoS)攻击是目前较常见的一类网络攻击行为,这类攻击以剥夺计算机和网络提供正常服务的能力为目的. 最常见的 DoS 攻击通过向攻击目标发送大量的攻击数据包来消耗目标主机或网络的资源,这类攻击通常被称为数据包洪泛攻击,例如 SYN 洪泛、UDP 洪泛、Smurf 攻击等[1].

分布式拒绝服务(Distributed Denial of Service,DDoS)攻击是洪泛式拒绝服务攻击一种更具威胁的演化版本,它利用因特网分布式连接的特点,通过控制分布在 Internet 上的计算机,共同产生大规模的数据包洪泛,对目标计算机或者网络进行攻击,例如 mstream, Stacheldraht, TFN2K 等[1].

收稿日期:2004-08-06;修改稿收到日期:2006-04-24. 本课题得到国家"八六三"高技术研究发展计划项目(2001AA144050,2003AA144050)资助.高 能,女,1976 年生,博士,目前主要从事拒绝服务攻击、蠕虫攻击、数据挖掘技术等方面的研究. E-mail:gaoneng@lois.cn. 冯登国,男,1965 年生,博士,研究员,博士生导师,目前主要从事信息与网络安全方面的研究与开发工作. 向 继,男,1976 年生,博士,目前主要从事信息与网络安全方面的研究与开发工作.

本文主要研究数据包洪泛 DoS 攻击的检测技术，目前检测这类攻击的方法主要还是基于数据包特征匹配的方法，Snort 等许多入侵检测系统都采用这种方法. 但是 DoS 攻击往往采用一些欺骗的手段，例如源 IP 地址欺骗和数据包属性欺骗，使得攻击数据包的表面特征与合法数据包相似或者完全相同，从而迷惑入侵检测系统[2]，使得它们无法进行有效的检测.

现有的入侵检测系统往往采用一种“亡羊补牢”的工作模式，即在主机或网络遭受某次攻击后，由安全专业人员花数小时来分析攻击数据包、总结攻击特征以及在 IDS 系统中配置特征规则. 这种工作模式存在两个缺点：首先，总结攻击特征主要靠安全专家手工完成，需要耗费大量的人力物力；其次，不能在攻击发生后很短的时间内检测出攻击，或者即使检测出攻击的发生，但是无法及时获得攻击特征进行防御，攻击造成的危害也就无法减少和消除.

本文提出了一种新的、基于数据挖掘的 DoS 攻击检测技术——DMDoSD，它利用数据挖掘算法中的关联算法和聚类算法分步处理数据，能够自动地、实时地、有效地检测 DoS 攻击. DMDoSD 不仅可以发出攻击报警，而且能够利用关联算法自动分析定位攻击特征，为 DoS 攻击的防御提供支持. 与现有的方案相比，DMDoSD 有以下几个优点：

(1)实时检测. 能够在攻击开始后 30 秒甚至更短的时间内发现攻击，有利于 DoS 攻击防御策略的快速布置.

(2)自动检测. 软硬件布置后，绝大部分的攻击检测工作可以自动完成，不需要人工参与. 自动检测同时也意味着能够更加快地完成检测.

(3)检测所有的数据包洪泛 DoS 攻击. 由于 DoS 攻击特征不固定，DMDoSD 不针对特定的 DoS 攻击类型，它利用关联算法自动挖掘出可疑的、感兴趣的数据包属性. 这种利用关联算法实现的自适应属性选择能够检测现有的所有数据包洪泛攻击类型及其变型，也同样适用于检测新的攻击类型.

(4)检测效率高. 只需进行一些简单的计算就能够进行检测，适合于在高速网络中布置.

(5)自动准确定位攻击特征. 能够自动地、准确地定位攻击数据包的特征.

本文第 2 节介绍 DMDoSD 的检测原理、检测过程以及特征发现；第 3 节给出利用 DMDoSD 检测 SYN 洪泛 DoS 攻击的实验情况；第 4 节介绍 DoS 攻击检测方面的一些相关研究工作和成果；第 5 节是结论部分.

2　DMDoSD 检测技术

2.1　检测原理

首先，DMDoSD 采用基于异常的检测方法. 异常检测的基本原理是分析正常数据，获得正常数据的模型，通过判断待检测数据是否偏离正常数据的模型来确定异常数据. 与异常检测相对的是误用检测方法，通常所说的数据包特征匹配方法就是一种误用检测方法. 与误用检测方法相比，异常检测有一个明显的优点就是能够检测出新的攻击类型.

其次，DMDoSD 采用了自适应检测模型产生方法[3]，该方法的基本工作原理如图 1 所示. 其中，数据采集模块依据数据收集策略采集网络中的数据(称为检测数据)，然后提交给特征提取模块. 特征提取模块将检测数据通过特定的算法转换成流量特征. 流量特征是对检测数据的高层抽象. 流量特征一方面输入到自适应模型产生模块用于产生检测模型，另一方面输入到攻击检测模块. 攻击检测模块根据流量模型判断该流量特征是否正常，从而产生检测结果. 自适应模型产生方法是在被保护网络环境中收集数据，产生最初的检测模型，并在积累到更多的数据后，通过具有自学习能力的模型产生算法自动产生新的检测模型. DMDoSD 利用自适应模型产生的这一特点来实现 DoS 攻击检测的自动性.

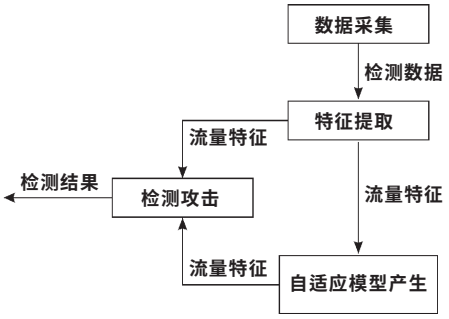

图 1　自适应检测模型生成示意图

基于异常检测的自适应模型产生方法需要特定算法的支持. 数据挖掘中的聚类算法是一类可选的方法[6]. 聚类算法的基本思想是将相似的数据事件聚集到一个聚类中，而且相同类型的数据应该距离很近，不同类型的数据之间距离很远[4]. 通过恰当地选择参数，聚类算法可以将正常的网络数据事件聚集在少数几个聚类中，我们称为正常聚类，而攻击数据事件将会远离正常聚类.

聚类算法的输入是表示网络流量的一些特征参数,包括目的地址、协议等等.这些参数选择的好坏直接影响到了利用聚类算法进行检测的性能.我们针对 DoS 攻击的特点,通过实验和理论推导了参数的选取方法,并引入关联算法进行参数的处理,提高了检测方法的性能.下面一节将详细描述 DMDoSD 的检测过程.

2.2 检测过程

DMDoSD 检测过程包括检测模型产生和攻击检测两个阶段:在检测模型产生阶段,网络上的原始数据被收集、处理并产生表示网络流量的特征参数(我们称之为流量特征),然后利用这些特征构造基于聚类的检测模型;在攻击检测阶段,类似地产生流量特征,然后利用检测模型进行检测.DMDoSD 中的检测算法采用了关联与聚类两种数据挖掘算法,下面结合检测模型和攻击检测进行描述.

2.2.1 检测模型

检测模型是攻击检测的基础,检测模型的产生分为流量特征产生和流量模型产生两个步骤,其产生过程如图 2 所示.

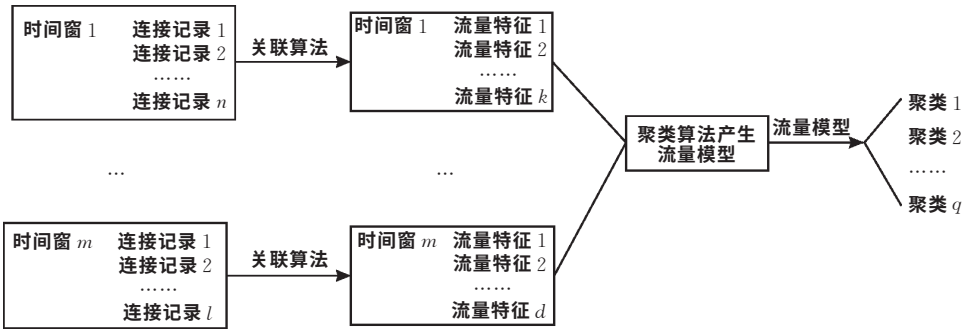

图 2 检测模型产生过程示意图

(1)流量特征产生

首先,DMDoSD 利用 tcpdump 等工具来收集网络上的原始数据包.通过预处理,将原始数据包转化为连接记录.连接记录包含了一个连接的发生时间、源地址、目的地址、协议类型、连接结束标志等属性,它不但可以描述 TCP 流量,也可以表示 UDP 和 ICMP 流量.DMDoSD 收集的连接记录内容如图 3 所示.

时间	源 IP 地址	目的 IP 地址	源端口	目的端口	协议类型	连接结束标志

图 3 连接记录的内容

DMDoSD 以时间窗为单位处理连接记录,即将网络数据根据发生的时间划分到一个个的时间窗内,将某段时间窗内的原始网络数据被恢复成连接记录.DMDoSD 在每个时间窗内进行一次检测,从而达到实时检测的效果.从 DoS 攻击特性考虑,将时间窗的大小定为 10s,这样既不会因为时间窗太小而导致漏检,也不会由于时间窗太大而增加检测的延迟.

接下来,DMDoSD 利用关联算法将连接记录转化为流量特征.DoS 攻击是一种群体网络行为,从单个 TCP 连接记录看都是正常的,但是在短时间内会出现大量的、类似的 TCP 连接,使用关联算法的目的就是捕获这种群体网络行为.关联算法是一种数据挖掘方法,能够发现数据记录中不同数据项之间的关联性.以网络数据为例,TCP 连接记录的每个属性就是数据项,关联算法可以计算出某个 TCP 连接记录的集合中所有属性组合出现的频率,比如"源 IP=192.168.0.133,目的端口=80"的组合,我们把那些超过某一数量的属性组合称为频繁项目集.

DMDoSD 使用 Apriori 关联算法[13] 对连接记录进行处理,通过规定属性组合为(服务类型,目的地址,TCP 连接结束标志)的频繁项目集为流量特征,确保产生有意义的规则,并引入两个表示数量的参数,如表 1 所示.属性组合中没有加入源 IP 地址是因为 DoS 攻击常常采用源 IP 地址欺骗达到分散攻击流量来源的效果,使得引入源 IP 地址后利用关联算法很难捕获 DoS 攻击的行为.

DMDoSD 使用 Apriori 关联算法将连接记录转化为流量特征,模型产生和攻击检测所处理的数据量大大减少,例如对于一个流量较大的网络来说,每个时间窗(10s)产生的连接记录可达上万个或者更多,而转化为流量特征后,只有几个或者十几个.数

据量的减少缩短了模型产生和进行检测的时间，同时增加了单位时间内网络数据处理量.

表 1　流量特征的内容

特征名称	特征含义
Service	服务类型(目的端口号)
dstIP	目的 IP 地址
Status	连接结束状态标志
count_conn	该时间窗内具有相同的 service、dstIP 和 status(即服务类型、目的 IP 和状态标志)的连接记录的个数
Count_total_conn	该时间窗内连接记录总个数

虽然关联算法可以获得对于 DoS 攻击流量一定的知识，例如在攻击发生时，连接结束状态是 S0(TCP 三次握手未建立的状态)的连接记录数量非常大，超过某个阈值，但是如何确定阈值、如何动态调整阈值等问题都无法解决，因而我们需要进一步利用聚类算法自适应的产生这些阈值，即需要下面的流量模型产生过程.

(2)流量模型产生

流量特征产生后，通过什么方式判断哪些流量特征是正常的，哪些流量特征异常的呢？DMDoSD 利用聚类算法挖计算出正常流量特征的聚类，再根据距离来判断异常. 聚类算法把流量特征视为向量，而服务类型等属性作为向量的分量. 聚类处理的结果就是大部分的正常数据组成一个或者几个大的数据集合，而攻击数据组成一些小的数据集合，它们与正常数据的距离较远，如图 4 所示. 我们把这些数据集合称为聚类，每一个聚类由该类的中心向量和半径表示.

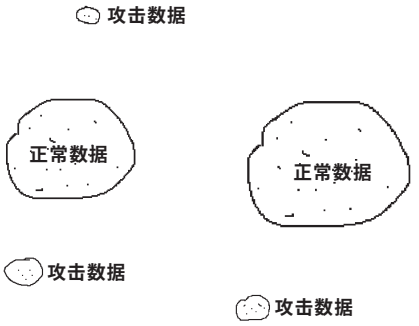

图 4　聚类模型示意图

DMDoSD 采用了 K-means 聚类算法来进行模型的产生. K-means 算法使用加权的欧几里德距离来度量流量特征之间的相似性：

$$d(i,j)=\sqrt{w_1|x_{i1}-x_{j1}|^2+w_2|x_{i2}-x_{j2}|^2+\cdots+w_p|x_{ip}-x_{jp}|^2}$$

给定一个包含 n 个数据的数据集，在给定聚类个数 k 时，K-means 算法将 n 个数据划分成 k 个子集，每个子集代表一个聚类，同一个聚类中的数据之间距离较近，而不同聚类的数据间距离较远. 每个聚类由其中心值来表示，通过计算聚类中所有数据的平均值可以得到它的中心值.

通常可以认为那些包含有大量数据的聚类是正常聚类(因为在实际的网络中 90%以上的流量都是正常的[6])，而那些只包含少量数据的聚类是异常的. 正常聚类组成了流量模型，这些流量模型被当作是检测模型发放到检测模块中.

聚类处理的结果并不是一个类似于 IDS 系统的匹配规则. 聚类的处理对象是大量的高维数据，并依据数据之间的距离度量它们之间的聚合状态，每个分量(属性)对距离都有贡献. 聚类处理的结果正是对训练数据聚合状态的表示，即流量模型，每个流量模型都是由中心点和半径表示.

模型的好坏与训练数据的质量有关. 存在虚警的原因主要是以前未见的网络行为被误认为攻击，这是在任何异常检测系统中都存在的. 但是正常行为不会像 DoS 攻击在短时间内产生大量的网络流量，即不会构成连续的异常，很容易就被识别. 当 DoS 攻击流量增大时，只要实施检测算法的平台能够足够快地处理网络流量，检测算法本身并不会导致漏警.

聚类算法能够自适应网络环境产生流量模型，因而可以布置在任何网络中.

2.2.2　攻击检测

DMDoSD 利用前面同样的方法收集单位时间窗内的网络数据，并利用关联算法产生若干流量特征. 接下来计算当前的流量特征向量与检测模型中的各个正常聚类的距离，如果该流量特征在所有的正常聚类之外，那么这个特征向量被认为是异常的，当某个特征向量在多个连续的时间窗都被标记为异常时发出报警.

通过以上的分析可以看出 DMDoSD 的模型产生算法和攻击检测方法不需要任何手工构造的训练数据，也不需要知道数据的性质，即不需要区分哪些是正常数据，哪些是攻击数据，这是由聚类算法本身的特点决定的. 而且，利用关联算法将数据集抽象为流量特征，不仅大大减少了聚类算法处理的数据量，克服了聚类算法运行速度慢的缺点，而且为聚类处理增加了有用的信息.

2.3　自动攻击特征发现

对于 DoS 攻击的防御来说，攻击特征的自动检

测是非常重要的一个需求,对于传统的入侵检测系统来说,攻击特征的分析和检测往往是人工完成的,这种分析一般至少要到攻击发生数小时候才有结果.而一般的 DoS 攻击会在 1 个小时或者几十分钟内造成很大的危害,这样即使攻击特征分析出来也为时已晚.相反地,如果能够自动地检测出攻击特征,在几分钟之内产生分析结果,并应用到网络防御系统中,DoS 攻击造成的危害将会大大降低.针对这种需求我们研究并提出了一种基于关联数据挖掘算法的攻击特征检测技术.

DMDoSD 检测到 DoS 攻击后,能够发出某个目标地址的某个目的端口遭受 DoS 攻击的报警,这说明 DoS 攻击数据包大都发往了这个目标地址,所以将所有发往这个目标地址的数据包收集起来进行进一步的分析,以发现攻击数据包的特征.

自动攻击特征发现技术需要两个前提,一个是在收集的所有数据包中,DoS 攻击数据包占了绝大多数,实际上在发生 DoS 攻击时,90% 以上的数据包都是攻击数据包;另一个是攻击数据包虽然会采用各种属性欺骗的手段以迷惑检测系统,但是还是会在某些属性上取相同的值或者呈现一定的规律,虽然在什么属性上相同以及该属性取何值是不固定的.

在上面这两个前提下,提取数据包的各种属性,并利用关联算法挖掘出属性组合的频繁项目集,那么这些频繁出现的属性组合很有可能代表了攻击数据包的特征,例如我们通过关联算法发现属性组合(协议类型 = TCP, 数据包长度 = 34, TTL 值 = 230)出现频繁,那么特征是(协议类型 = TCP, 数据包长度 = 34, TTL 值 = 230)就很有可能是 DoS 攻击数据包.

DMDoSD 提取如下的数据包属性作为关联算法的输入:协议类型、数据包长度、源地址、目的地址、TTL 值、TCP 标志、源端口、目标端口、IP 标识(identification)、TCP 序列号、TCP 窗口号等.关联算法分析的结果是若干属性组合的频繁项目集,这些属性组合很有可能代表了 DoS 攻击数据包的特征,但是在没有攻击时,也有可能产生一些属性组合的频繁项目集.为了保证获得的属性组合真正地代表攻击数据包的特征,我们在没有攻击时使用同样的方法获得那些频繁出现的属性组合,在发生攻击时,从分析得到的那些属性组合中去除这些网络正常情况下获得的属性组合,那么剩下的属性组合基本上可以认为是代表了 DoS 攻击数据包的特征.

3 攻击检测实验

为了验证检测算法的有效性,对 DMDoSD 进行了 DoS 攻击的实验,对检测算法的性能进行了测试.通过实验发现,DMDoSD 对数据包洪泛的 DoS 攻击有很高的检测率,而且误警率很低.

实验中使用的网络数据包括两个部分:(1)背景数据,即正常的网络流量;(2)攻击数据,即 DoS 攻击流量.我们采用的背景数据是美国林肯实验室公开的网络数据[7],它们是从一个模拟的硬件网络环境下采集的,具有一定的代表性,被国内外的研究机构广泛地用来进行入侵检测的实验.但是其中包含的攻击特别是 DoS 攻击都比较陈旧,不具代表性,所以只作为背景数据.攻击数据利用 Syn flooder 工具产生,根据需要产生不同攻击速率和不同时间长度的攻击数据.

实验利用林肯实验室第一个星期的数据(共 213205 条连接记录)作为背景数据产生检测模型,产生的模型包含 7 个正常数据的聚类,这 7 个聚类包含了 98% 的网络数据;然后利用 Syn flooder 工具随机产生 7 次 SYN 洪泛攻击,每次采用不同的攻击速率和时间长度,而且都采用随机方法产生攻击数据包的源 IP 地址;接下来将产生的攻击数据随机插入到林肯实验室的第二个星期的数据(共 216830 条连接记录)中产生待检测数据;最后将待检测数据送入检测器进行检测,产生的检测结果如表 2 所示.

表 2 SYN 洪泛攻击检测性能

ID	攻击速率(次数/s)	持续时间(s)	检测时间(s)	检测率(%)	误警率(%)
1	5	20	0	0	0
2	10	32	0	0	0
3	15	31	0	0	0
4	30	30	6	100	0
5	100	30	10	100	0
6	500	22	14	100	0
7	3000	23	8	100	0

通过分析表 2 中的数据,可以得出以下的一些结论:

(1)检测算法对攻击速率比较高的 DoS 攻击检测率高,但是检测不出速率低于 30 次/s 的 DoS 攻击.实际网络环境中洪泛 DoS 攻击的攻击速率远远高于 30 次/s,所以该检测算法对现实中的 DoS 攻击是有效的.

(2)检测算法能够在 20s(两个时间窗)以内检测出 DoS 攻击,具有很好的实时性.

(3)检测算法的误警率几乎为 0，不会像目前的入侵检测系统那样产生许多误报. 当然这与背景数据的情况有关，我们今后将在不同的网络环境中收集背景数据，以进行进一步的验证.

DMDoSD 不适于检测慢速 DoS 攻击的原因是：慢速 DoS 攻击的网络行为已经非常接近正常流量的网络行为，例如慢速 Syn 洪泛如果以 15packet/s 的速率进行攻击，正常网络流量也可能出现如此频率的 TCP 连接失败率. 慢速 DoS 攻击检测可以采用基于时间窗的周期性分析方法进行补充. 慢速 DoS 攻击在发生时通常会持续一段时间，周期性分析可以捕获这种持续的、低速的 DoS 攻击. 低速 DoS 攻击是无法在短时间内对目标构成破坏的，特别是具有强大处理能力的服务器或者大带宽网络. 因而在本文中没有专门论述对低速 DoS 攻击的检测.

与现有的基于数据挖掘算法的 DoS 检测方法相比，DMDoSD 最大的优势就是适合实时检测. 哥伦比亚大学的 Portnoy 等人提出了利用基于距离的聚类算法进行入侵检测，但是这种单纯基于聚类算法的检测方法需要大量的运算，也不是专门为检测 DoS 攻击的，其对平均攻击检测率在 40%～55%，虚警率在 1.3%～2.3%之间. 美国明尼苏达大学 Aleksandar[14] 在利用 DARPA 1998 的数据对于多种数据挖掘算法包括 k 次 Nearest Neighbor、Nearest Neighbor、基于 Mahalanobis 聚类的 Outlier 检测、基于密度的局部 Outlier 检测以及无监督的 Support Vector Machines(SVM)在内的 5 种异常检测算法进行了比较. 实验发现除了 Mahalanobis 算法较低，其余算法对于 DoS 攻击的检测率都在 70%左右.

但是，由于 DoS 攻击检测和防御与训练数据和网络环境有很大的关系，所以我们认为直接与其他研究成果进行检测率的比较并不科学.

与我们方法比较相近的研究还有美国明尼苏达大学提出的“Minnesota Intrusion Detection System”简称 MINDS 方案，在该方案也采用了数据挖掘算法，但是它的目的是设计和实现一种通用的 IDS 系统，也没有专门针对 DoS 攻击的检测更像 Snort.

4　相关研究

DoS 攻击实时检测是目前国内外在网络安全方面的一个研究热点，各类研究机构都提出了自己的方案.

美国哥伦比亚大学的 Mohiuddin 等人提出了一种基于数据挖掘的 Dude 方案[8]，它主要利用贝叶斯分类挖掘算法来检测攻击. 首先，经过标记的训练数据(正常数据和攻击数据)由汇总器处理后，划分为一些数据块. 特征提取器提取出这些数据块的特征，模型生成器利用贝叶斯分类算法为这些特征产生一组规则，实际上就是这些特征的阈值. 在检测时，首先提取出网络数据块的特征值，然后与流量模型中的阈值比较，如果某一特征值超出了相应的阈值，那么就认为是异常.

与之类似的方案还有 ADAM[9]、PHAD、NIDES 等，这些方案的缺点是需要大量人工构造的训练数据，而这些数据只有从真实的应用环境中获得才有价值，而在实际应用中，获得这些训练数据是相当困难的，同时它们也不适于检测新的攻击类型.

针对这种情况，国外研究者开始研究无需训练数据的检测方案，这类方案主要是利用聚类算法处理训练数据，找到那些偏离正常数据比较远的数据，如何度量这种偏离是其中最关键的问题，由此产生了许多方法：基于距离的、基于密度的和基于模型的方案.

哥伦比亚大学的 Portnoy 等人提出了利用基于距离的聚类算法[5]进行攻击检测，它们从原始网络数据的连接记录中构造聚类，并为这些聚类打上标记表明它们是正常还是异常，并利用这些标记的聚类来对后续的网络数据进行分类. 这种方案直接利用连接记录作为聚类算法的输入，所以误警率比较高，而且不适于进行实时检测.

美国明尼苏达大学提出的 MINDS 方案[6]是此类方案中较有代表性的一个. 它的异常检测模块利用基于密度的方法，计算每个连接记录的 LOF 值，并为每个网络连接记录赋予一个得分，该得分反映了该连接记录偏离正常网络数据的程度.

前面介绍的主要是基于数据挖掘技术的 DoS 攻击检测方法，此外还有很多基于统计方法的检测系统，例如 SPADE 系统[10]，它检查网络连接的属性组合概率，概率越低，表明其为异常的可能性越大. SPADE 存在的一个明显的问题就是虚警概率较高，因为合法流量中的许多属性组合并不是频繁发生的，其概率较低. 特别是对于那些由内部网络向外发起的连接，由于网络外部 IP 地址数量比网络内部 IP 地址数量要多得多，实际的训练数据就不能完全准确地刻画某些组合属性的统计分布特性.

此外澳大利亚墨尔本大学 Peng 等人提出了一

种利用源 IP 地址统计分布的检测方法[11]，它以源 IP 地址为统计对象，将流量在地址空间的分布进行记录. 正常情况下，在某一时刻，只有有限多的 IP 地址有进入流量，而当 DDoS 攻击发生时，就会出现许多 IP 地址同时请求连接、建立连接的情况. 这种方案的缺点是误警率较高，对于一个大型的网站来说，很多个 IP 地址同时访问往往是正常的.

5 结 论

本文针对 DoS 攻击难于检测和难于防御的特点，提出了利用数据挖掘算法实现 DoS 攻击的实时检测以及攻击数据包特征的实时发现技术——DMDoSD.

数据挖掘算法通常需要进行复杂的计算，所以被认为不适合进行在线的实时检测[12]，为了解决这一问题，我们利用关联算法从网络流量中提取有用的流量特征并大大减少聚类算法处理的数据量，从而在不影响检测性能的情况下减少了检测算法对计算能力的需求.

如果不能够实时地发现 DoS 攻击数据包的特征，攻击检测算法的意义就失去了一大半. 通常这种工作由安全专家手工完成，需要大量的时间和复杂的过程. DMDoSD 同样利用数据挖掘中的关联算法自动地发现攻击数据包的特征，从而也解决了这一问题.

与现有的其它技术相比，该技术具有很强的实用性. 目前使用该技术的原型系统已经实现，并申请了国家发明专利(专利申请号：02155382.3). 在实现过程中，将数据收集、实时检测和模型生成三个相对独立的工作分担到三个不同的设备上进行，提高了处理的并行性和实时性，完全满足百兆级网络对处理能力的需求.

参 考 文 献

1 Skoudies. Counter Hack. Beijing: China Machine Press, 2002 (in Chinese)
(Skoudis Ed. 反击黑客. 北京：机械工业出版社, 2002)

2 CERT/CC Coordination Center. Trends in Denial of Service Attack Technology. October 2001

3 Honig A., Howard A., Eskin E., Stolfo S.. Adaptive model generation: An architecture for the deployment of data mining-based intrusion detection systems. Data Mining for Security Applications, Kluwer, 2002

4 Han Jia-Wei, Kamber M.. Data Mining: Concepts and Techniques. Beijing: Higher Education Press, 2001(in Chinese)
(韩家炜等. 数据挖掘——概念与技术. 北京：高等教育出版社, 2001)

5 Portnoy L., Eskin E., Stolfo S. J.. Intrusion detection with unlabeled data using clustering. In: Proceedings of the ACM CSS Workshop on Data Mining Applied to Security (DMSA, 2001), Philadelphia, PA, 2001

6 Ertoz L., Eilertson E., Lazarevic A., Tan P., Dokas P., Srivastava J., Kumar. Detection and summarization of novel network attacks using data mining. Technical Report, 2003

7 The 1999 DARPA Intrusion Detection Evaluation Data set, Information Systems Technology Group of MIT Lincoln Laboratory, http://www.ll.mit.edu/IST/ideval/data/data_index.html

8 Mohiuddin S., Hershkop S., Bhan R., Stofo S.. Defending against a large scale denial-of-service attack. In: Proceedings of the IEEE. Workshop on Information Assurance and Security, New York, 2002

9 Barbara D.. ADAM: Detecting intrusions by data mining. In: Proceedings of the 2001 IEEE. Workshop on Information Assurance and Security, 2001

10 Staniford S., Hoagland J., McAlerney J.. Practical automated detection of stealthy portscans. Journal of Computer Security, 2002, 10(1/2): 105～136

11 Tao Peng, Christopher Leckie, Kotagiri Ramamohanarao. Detecting distributed denial of service attacks using source ip address monitoring, draft, 2002

12 Brugger S. T.. Data mining methods for network intrusion detection. Technique Report, UC davis, 2004

13 Christian Borgelt. The apriori program source code. http://fuzzy.cs.uni-magdeburg.de/～borgelt/, 2002

14 Lazarevic A., Ertoz L., Ozgur A., Srivastava J., Kumar V.. A comparative study of anomaly detection schemes in network intrusion detection. In: Proceedings of the 3rd SIAM Conference on Data Mining, San Francisco, 2003

GAO Neng, born in 1976, Ph. D.. Her research interests include denial of service, worm attack and data mining technique, etc.

FENG Deng-Guo, born in 1965, professor, Ph. D. supervisor. He mainly engaged in the research and development of information and network security.

XIANG Ji, born in 1976, Ph. D.. He mainly engaged in the research and development of information and network security.

Background

This paper is based on the research results of two projects, "Robust Gateway Technology" and "Secure Gateway Technology", which are supported by the National High-Tech Research and Development Program of China (grant No. 2001AA144050 and 2003AA144050). These two projects are to develop a novel gateway-based system to protect application or enterprise networks from DoS and DDoS attacks. Efficient system architecture and attack detection algorithms are emphasized in these two projects.

Real-time DoS attack detection gains considerable attentions in the network security domain, and many valuable projects are brought forward by different research institutes. Suhail Mohiuddin from the University of Colombia describes a data mining based method called Dude which introduces a Naïve Bayes rule-based classifier to learn an attack model using training data. There are also some similar schemas such as ADAM, PHAD and NIDES. These schemas are limited in that training data must be collected from real networks and manually flagged as normal or abnormal. Additionally, these methods cannot detect novel attack types.

To solve the above problems some researchers have focused on unsupervised data mining methods. Data mining techniques are introduced into intrusion detection systems to improve their effectiveness and efficiency. Among all the data mining-based intrusion detection technologies, cluster-based intrusion detection is the most promising one because of its ability to detect new attacks. Single-linkage and k-means cluster algorithms are widely utilized to solve anomaly detection problems. Single-linkage algorithm has higher efficiency but lower precision. K-means algorithm exhibits better performance but needs to scan traffic data multi-times thus cannot meet real-time requirement. Huge amount of traffic packets are the key element effecting processing efficiency, especially during DDoS floods. Previous per-packet based processing method hampers speed improving.

The authors put forward a new concept of traffic feature, which summaries collective behaviors of large TCP connections in a same time-window. An association algorithm such as Apriori, can be utilized to achieve traffic features, which are directly processed by a cluster algorithm. With association algorithm and cluster algorithm cooperation, a good balance between efficiency and performance is acquired.

The method explained in this paper is mainly the DoS attack detection method of these two projects. Furthermore, a patent is applied for it named as "A Robust Gateway System and Method", grant No. 02155382.3. At present, the prototype system is servicing in the lab's local network and some real-time DoS attacks test results are collected.