

用 C4.5 算法对局域网数据报进行行为分类

吴楠, 宋方敏

(南京大学软件新技术国家重点实验室, 江苏南京 210093)

摘要: 局域网传输的数据报中携带大量与数据包相关的信息, 这些信息在一定程度上反映了数据报的行为。对数据报行为进行分类可为局域网上的网络入侵检测提供重要依据。文中提出使用 C4.5 决策树分类算法对局域网数据报进行行为分类, 并与以往常用的几种分类算法进行了比较。实验表明, C4.5 算法对于该问题无论在分类效率还是在分类正确性方面均有很大的优势。

关键词: C4.5 算法; 数据挖掘; 局域网数据报; 网络入侵检测; 分类; 决策树

中图分类号: TP311

文献标识码: A

文章编号: 1673-629X(2006)07-0001-03

Classify Datagram on LANs by Using C4.5 Classification Algorithm

WU Nan, SONG Fang-min

(State Key Laboratory of Novel Software Technology, Nanjing University, Nanjing 210093, China)

Abstract: A mass of package-related information is carried by datagram on LANs (local area networks). This information reflects the action and behavior of the datagram. Classify the datagram on LANs can provide some evidences of NID (network intrusion detect). A decision tree-based classification algorithm by using C4.5 algorithm is introduced in this paper. Compared with other classification algorithms that usually used, this method shows distinctly advantages in both efficiency and precision in classifying datagram on LANs.

Key words: C4.5 algorithm; data mining; LAN datagram; network intrusion detect; classification; decision tree

0 引言

计算机局域网上绝大部分信息的传递都通过数据报完成。由于局域网的相对开放性, 可能有部分不良行为的数据报侵入到局域网中并对局域网内计算机中的信息产生危害, 这种危害包括数据的丢失和改变、降低联网计算机的工作性能、造成局域网相关服务的中断以及机密信息的泄漏等严重问题。为了避免相关危害的产生, 需要监控局域网上所有的数据报并识别出可能有不良行为的数据报的类型, 然后对相关的的数据报报文进行检测。这便是称之为“网络入侵检测(Network Intrusion Detect, NID)”的过程。NID 首先关心的问题是如何发现可能存在不良行为的数据报, 这一般是根据位于数据报报头的各个包属性(Package Attribution)将数据报按行为进行分类, 然后找出可能具有不良行为特性的数据报来完成。

为了得到包属性到数据报行为类型的映射关系, 需要总结出一系列由包属性到数据报行为类型的映射规则。由于局域网上数据报的流量特别巨大, 依靠人工找到相关规则和进行分类是不可能的, 因此需要用数据挖掘的方法对上述映射规则进行提取并利用这些映射规则对新的数

据报进行分类。

1 局域网数据报及其行为类型

数据报是携带各种信息的, 独立从数据源行走到终点的数据包。局域网上信息的交换绝大多数通过数据报传送。一个完整的数据一般包含 41 个包属性(见表 1, 根据协议的不同, 不同类型的数据报可能具有不同的属性):

表 1 局域网内数据报的包属性

属性编号	属性名	属性值数据类型
1	duration	continuous
2	protocol	symbolic
3	service	symbolic
4	flag	symbolic
...
40	d-h-error-rate	continuous
41	d-h-srv-error-rate	continuous

其中, 属性值数据类型中标记“continuous”的为连续型数据类型, 如整数、浮点等; 标记“symbolic”的为离散型数据类型, 如字符、字符串等类型。

数据报按照其行为可分为正常的数据报与非正常的数据报。正常的数据报的行为具有负责传递有用数据、控制命令等合法用途, 而非正常数据报就有可能是以攻击局域网内计算机或者破坏局域网服务为目的的。非正常的数据报一般可以分为以下几类:

* DoS(Denial of Service)或 DDoS(Distributed Denial

收稿日期: 2005-12-29

作者简介: 吴楠(1981-), 男, 山东济南人, 硕士研究生, 研究领域为软件新技术与软件方法学、信息检索与获取等; 宋方敏, 教授, 博士生导师, 研究方向为理论计算机科学。

of Service)攻击^[1]:拒绝服务攻击与分布式拒绝服务攻击:利用 TCP 协议缺陷,发送大量伪造的 TCP 连接请求,从而使得被攻击方资源耗尽(CPU 满负荷或内存不足)的攻击方式。例如,在短时间内向服务器端发送极大量的同步报文(SYN-gram),以同步报文泛洪攻击(SYN flood)的方式造成服务器端瘫痪。

* R2L 攻击:远程计算机对局域网内的计算机发送未经认证的数据报报文。未经认证的报文有可能是对局域网内计算机密码的猜测报文,有可能对其安全性产生危害。

* U2R 攻击:对局域网内计算机最高访问权限(Root Privilege)的未经认证的访问。这种未经认证的访问攻击可能来源于局域网内或网外的缓冲区溢出攻击,通过一定的程序设计技巧造成网内服务器计算机缓冲区溢出,从而获得非法的最高访问权限,这种攻击会对网内计算机产生相当巨大的危害。

* Probing 攻击:通过监视或者扫描网内服务器或联网计算机,获得网内计算机可能的漏洞,并根据该漏洞设计攻击网内计算机的程序。例如,使用监视程序扫描网内计算机的端口以获得端口是否打开的信息,然后从可能有漏洞的端口尝试向该计算机发起攻击。

2 对局域网数据报行为进行分类

任何事物的行为都有可能在它的特征上侧面地表现出来,数据报也是一样。根据数据报的 41 个包属性值,可以从中发现一些有用的、与数据报行为相关联的规则。利用这些关联规则,可以将新的、行为未知的数据报进行行为分类。

2.1 决策树分类方法

决策树(Decision Tree)是一种类似于流程图的树结构,其中每个内部节点表示一个属性上的测试,每个分支代表一个测试的输出(Yes/No),每个叶结点代表划分的类或者类的分布。最顶端节点为根节点^[2]。

为了对未知样本进行分类,样本的属性值在决策树上测试。路径由根到存放该样本测试的叶结点。判定树可以很容易地转换为分类规则^[3]。

要进行决策树分类,首先必须构造一棵决策树。下面介绍决策树构造的 C4.5 算法:

C4.5 算法是 J. R. Quinlan 于 1993 年提出的一种对 ID3 的改进算法^[4]。C4.5 算法克服了 ID3 在应用中的一些不足,有一些独特的优点^[5]:

- * 用信息增益率进行判定属性的选择,克服了用信息增益选择判定属性时偏向选择取值较多的属性的不足;
- * 在决策树的构造过程中或者构造完成之后进行决策树的剪枝;
- * 能够在算法历程中完成对连续属性的离散化处理,即输入数据中可以包含连续值属性;
- * 能够对不完整的数据进行处理,例如能对未知的

属性值进行处理;

- * 增强了数据的适应性;
- * 可以处理具有不同代价的属性;
- * C4.5 算法最终可以形成产生式规则;
- * 降低了计算复杂度,增强了计算的效率。

C4.5 算法对于 ID3 算法的重要改进是使用信息增益率(Information Gain Ratio)来选择属性。理论和实验表明,采用信息增益率比采用信息增益更好,主要是克服了 ID3 方法选择偏向取值多的属性。

C4.5 算法还针对连续值属性的数据进行了处理,这弥补了 ID3 算法只能处理离散值属性数据的缺陷。这使得 C4.5 算法应用在同时具有连续值和离散值属性的网络数据报数据中对其进行行为分类具有极大的方便。

2.2 用 C4.5 决策树生成算法对局域网数据报进行行为分类

由于模拟局域网数据报数量非常巨大,而且在其 41 个属性中有 34 个为连续值属性,因此,C4.5 算法对于从数据报中挖掘包属性与数据报行为之间的关联规则是合适的。

图 1 显示了用 C4.5 算法对局域网数据报进行行为分类的流程图。

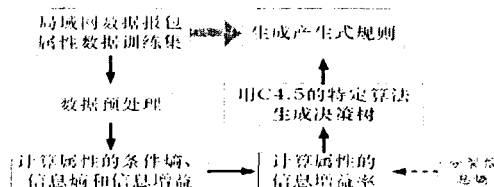


图 1 基于 C4.5 算法的局域网数据报行为分类流程图

图中,数据预处理的过程一般指对原始数据的净化和维度的约减,由于应用的特殊性,对原始数据(即输入的局域网数据报包属性数据)的噪声处理具有很大的难度,同样地,为了保证产生规则的准确性,在运行时间允许的情况下也不对 41 个输入属性当中的任何属性进行约减。由于算法不需要输入离散值得数据,因此也不需要连续数据进行离散化。对输入数据进行的预处理仅是建立数据链表以及将离散的属性进行数值标记。

进行数据预处理后,计算所有属性的条件熵、信息熵和信息增益,并根据这些数值和分裂信息熵计算出属性的信息增益率。最后完成决策树的构造。这里特定的算法主要指对连续值属性数据的划分。

在完成局域网数据报包属性-数据报行为的规则决策树的构建之后,可以使用 IF-THEN 语句完成决策的产生式规则。产生式规则列举了包含每个属性的 IF-THEN 语句。

最后,根据产生式规则集,对每一个新的数据报包属性的数据,就可以在仅进行一遍扫描的情况下完成对该数据报的行为分类。

2.3 用 C4.5 算法生成局域网数据报行为分类决策树

各种不同的分类算法相互差异性很大,选择决策树算

法特别是C4.5算法对局域网数据报行为进行分类是出于如下考虑:

(1)数据报包属性中所有的值均为明确的,利用决策树很容易产生出IF-THEN的产生式决策规则;

(2)数据报包属性中超过80%的数据是连续值类型的,若使用诸如贝叶斯分类器(Bayesian Classifier)等方法进行分类必须首先对这些连续值属性进行离散化,而离散化的方法随意性很大,效果难以保证。C4.5算法在ID3算法的基础上利用信息增益率进行连续值属性的阈值分化,分化效果良好且稳定。

(3)局域网上的数据流速度快,而且不可能重复出现已经出现过的流数据,因此,这对分类器的速度要求是很高的,有的分类器由于要进行大量复杂的运算,处理速度跟不上网内数据的吞吐量,需要开很大的缓冲器存放待处理的数据,造成了效率的下降和资源的浪费;而有的算法需要对数据流扫描多趟,这种算法就完全不适合于这种应用背景。C4.5算法生成的是一组简单的IF-THEN产生式规则,利用这些规则,仅对数据进行一次扫描就完全拥有足够的信息完成分类工作。

3 用C4.5算法对局域网数据报行为进行分类的结果

对于验证该训练器,笔者进行了实验。首先在一个包含247 010个数据报包属性的训练数据集中随机挑选1/3的数据进行训练,然后对训练的结果(决策树)用另外2/3的数据进行检验,得到分类器对这组数据的分类正确率为99.8043%。

最后对247 010个数据集用C4.5算法产生了完整决策树并进行了剪枝。没有对决策树应用进行繁琐的剪枝算法(如C4.5-J48算法的剪枝),而是让程序自然地将支持度较低的分枝去掉。得到的剪枝后的决策树可以用IF-THEN产生式规则描述如下(类C伪代码片断):

```
if (count > 64.0)
{
    if (dst_b <= 0.0)
    {
        if (diff_srv_rate <= 0.29) strcpy(dgclass, "dos");
        else strcpy(dgclass, "probe");
    }
    else strcpy(dgclass, "normal");
}
else if (id23_count < 64.0)
{
    if (compromised > 0.0)
    {
        if (src_b <= 22532.0)
        {
            if (d_h_same_src_port_rate <= 0.21) strcpy(dgclass,
"normal");
```

```
else strcpy(dgclass, "u2r");
}
else strcpy(dgclass, "dos");
}
...
}
```

4 结果分析与实验结论

用本算法和几种不同的分类方法对另外提取出来的5个与训练集独立的数据集进行了分类实验,每个实验数据集包括651 300个数据,结果如图2和图3所示。易见,C4.5算法在训练速度、分类速度和准确度上综合表现良好,而且对不同数据分布形式的数据集表现稳定,对于基于数据挖掘的入侵检测来说是一种较好的分类算法。

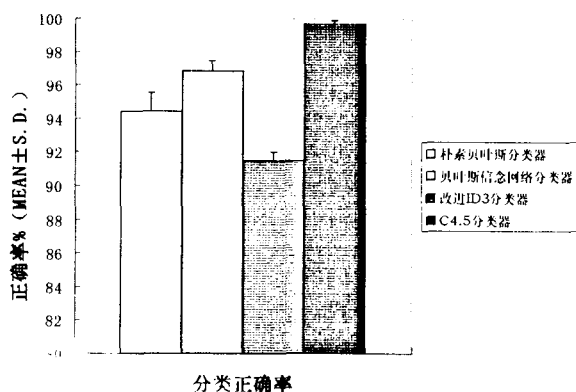


图2 4种分类器的分类正确率比较

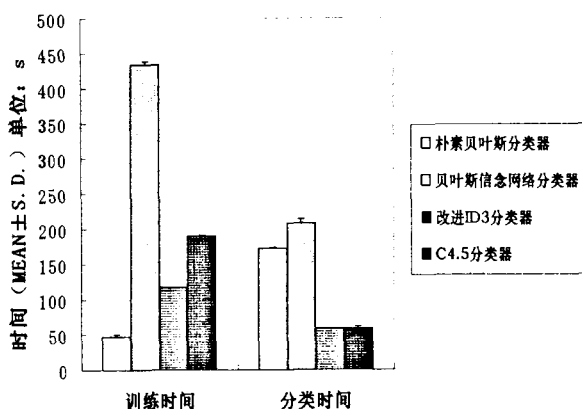


图3 4种分类器的训练和分类时间比较

参考文献:

- [1] 沈芳阳、李振坤、柳正青,等.DDoS攻击及其防范策略[J].微机发展,2003,13(9):46-48.
- [2] Quinlan J R. Induction of decision trees[J]. Machine Learning, 1986, 1: 81-106.
- [3] Han Jiawei, Kamber M. Data Mining: Concepts and Techniques[M]. [s.l.]: Morgan Kaufmann Publishers, 2001.
- [4] Quinlan J R. C4.5: Programs for Machine Learning[M]. San Mateo, CA: Morgan Kaufmann, 1993.
- [5] 陈文伟、黄金才、赵新昱.数据挖掘技术[M].北京:北京工业大学出版社,2002.