基于协议分析和决策树的入侵检测研究

胡琼凯，黄建华
（华东理工大学 信息科学与工程系，上海 200237）

摘 要：入侵检测技术已经成为网络安全领域的热点。针对现有入侵检测系统的不足，尤其是检测效率低、不能有效检测未知入侵和新的攻击行为等，在深入研究协议分析的基础上，通过引入数据挖掘技术中成熟的决策树分类算法，设计出了一种融合协议分析和决策树技术的动态、可扩展的入侵检测模型。同时，建设性地提出了分类组织检测规则的思想，检测过程引入了正则表达式匹配技术，有效地实现了对检测模型的优化。实验结果表明，该方法能够有效地提高检测速度、降低正误报率，具有较好的理论意义和实际意义。

关键词：决策树；协议分析；入侵检测；正则表达式
中图分类号：TP393.08 **文献标识码**：A **文章编号**：1673-629X(2009)06-0179-03

Intrusion Detection Based on Protocol Analysis and Decision Tree

HU Qiong-kai, HUANG Jian-hua
(Dept, of Information Science and Engineering, East China University of Science and Technology, Shanghai 200237, China)

Abstract: Intrusion detection technology has become the hotspot of network security. Considering the deficiencies especially like low efficiency, unable to detect those unknown and new behavior etc., in current intrusion detection system, this research, based on protocol analysis, presented a better implementation for intrusion detection by adopting the decision tree theory of data mining technology. Meanwhile, to optimize the dectecting model, it innovatively brought up an idea for refactoring detection patterns by categories, and also introduced the regular expression theory into the detecting practice. Research shows that this method has higher performance and less false positives, and has good theoretical and practical significance.

Key words: decision tree; protocol analysis; intrusion detection; regular expression

0 引 言

目前，绝大多数 IDS(入侵检测系统)的检测机制还停留在：基本的数据包捕获加以非智能模式匹配与特征搜索技术来探测攻击。然而，简单的特征模式匹配技术[1]存在着两个最根本的缺陷：匹配算法计算量大、特征匹配误报率较高。

协议分析技术[2]充分利用网络协议的高度规则性，对数据包中相应位置的协议信息进行解析，只检测对入侵检测有用的内容。能够智能地“理解”网络协议，确保特征串的实际意义被正确解析，有效地降低了误报率；通过引导搜索数据包明确特定的部分而不是整个有效载荷，快速探测攻击的所在，避免了简单模式匹配所做的大量无用功，提高入侵检测的效率。

入侵检测可以被看作是对网络数据正常或入侵的预测，即通过一个分类模型将网络数据分为入侵和正常两类。文中将数据挖掘技术中的决策树[3]算法与协议分析技术结合运用于入侵检测[4]。该方法对于提高检测效率和降低误判率有很好的作用。

收稿日期：2008-10-05；**修回日期**：2009-01-07
作者简介：胡琼凯(1984-)，男，硕士研究生，研究方向为网络入侵检测；黄建华，硕士生导师，教授，研究方向为网络与信息安全。

1 Ethereal 协议分析

Ethereal 是当前较为流行的一种计算机网络调试和数据包嗅探软件。Ethereal 的功能基本类似于 TCPdump，但它可以同时支持 Linux 和 Windows 平台，并且是跨平台、开源的。目前 Ethereal 已经支持 500 多种协议的解析，功能可以和商业的网络协议分析系统媲美[5]。

借助 Ethereal 协议分析器，可以方便地对捕获到

表 1 决策树训练数据样例表

目标 IP	协议	服务	请求 URL	…	类别
219.133.48.118	TCP	HTTP	/audio.dll	…	I
202.118.66.251	TCP	HTTP	/bin/ps	…	I

```
1809 18.661694 61.172.201.138    192.168.1.199     HTTP   HTTP/1.0 200 OK (application/>
1799 18.548860 192.168.1.199     61.172.201.138    HTTP   GET /js/2007live.js HTTP/1.1
1790 18.509990 192.168.1.199     125.86.82.167     HTTP   Continuation or non-HTTP traff
+ Internet Protocol, Src Addr: 192.168.1.199 (192.168.1.199), Dst Addr: 61.172.201.138 (61.
+ Transmission Control Protocol, Src Port: 1420 (1420), Dst Port: http (80), Seq: 1, Ack: 1,
- Hypertext Transfer Protocol
  + GET /js/2007live.js HTTP/1.1\r\n
    Accept: */*\r\n
    Referer: http://sports.sina.com.cn/nba/\r\n
    Accept-Language: zh-cn\r\n
    Accept-Encoding: gzip, deflate\r\n
    If-Modified-Since: Mon, 22 Dec 2008 01:00:02 GMT; length=2034\r\n
    User-Agent: Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1; SV1; Mozilla/4.0(Compatit
    Host: nba.sports.sina.com.cn\r\n
    Connection: Keep-Alive\r\n
```

图 1 Ethereal 数据包协议分析结果

的网络数据包(见图 1)进行处理,形成以属性 - 值形式表示的数据(见表 1),用于生成入侵检测决策树。

2 入侵检测决策树

2.1 决策树算法思想

决策树算法基本思想为:假设 R 为训练样本集,每一个样本均有若干个属性。从样本集根节点出发,根据对属性值的测试结果逐渐分裂产生分支节点,直到产生一颗完整的树。要为 R 构造决策树必须首先确定属性测试选择标准,选出当前检测属性。常用的属性选择标准有信息增益、信息增益率和 gain 等。按属性选择标准并按当前属性的 n 个值将 R 分裂为 n 个子集。若第 i 个子集 R_i 含有的所有记录的类别标签一致,该节点就成为决策树的叶节点,停止分裂并以所属分类标记该叶节点;而对不满足此条件的 R 的其他子集则按照上述方法继续分裂,直到所有子集所含记录均属于一个类别为止[6]。

2.2 决策树经典算法

ID3 算法的基本核心是采用贪心算法[7],它采用自上而下、分而治之的递归方式来构造一棵决策树,但因其选择具有最高信息增益的属性作为测试属性,只能处理离散属性。C4.5 算法作了改进,选择具有最高信息增益率的属性作为测试属性,能够有效地处理连续属性。文中使用 C4.5 算法生成入侵检测决策树[8]。

2.3 入侵检测决策树的生成

a)建立数据分类和属性文件。数据分类包含入侵(I)和正常(N)两类。属性文件用于存放属性名、属性类型以及属性对应的所有可能取值。

b)准备训练数据。由 Ethereal 对捕获到的网络数据包进行协议分析预处理,形成属性 - 值形式的训练数据。

c)使用 C4.5 算法,根据信息增益率选择测试属性创建决策树。例如,由表 1 中的测试样例数据生成图 2 所示的入侵决策决策树。

2.4 决策树入侵检测过程

检测引擎通过遍历入侵检测决策树来实现检测。网络数据经过协议分析处理后,转换成与训练数据相同的格式。将属性 - 值型记录中与根节点测试属性对应的属性值提取出来,并将此值与分支值比较。如果找不到匹配的分支,则该记录检测结束,此记录的分类为该节点的默认最佳分类。如果默认分类为入侵,则立即报警。如果此属性值与某个分支值匹配,该分支将此记录指向下一个子节点,再对下一个节点属性进行比较。依此往复,直至到达叶节点为止。叶节点所标记的分类为该记录的分类[9]。因此,对一条记录进行检测,它所走过的路程就是决策树中的一条从根节点到叶节点的路径。与必须一条条进行匹配的基于规则的简单模式匹配相比较,节省了许多冗余的比较次数。

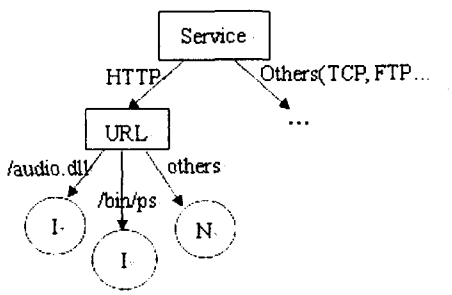

图 2 由测试样例生成的入侵检测决策树

3 分类组织检测规则

鉴于简单模式匹配集中处理匹配规则的方法,受益于决策树思想的启发,文中建设性地提出对检测规则进行分类组织,即在脱机方式训练学习数据、构建决策树的过程中,根据决策树中包含的不同的 Service 属性值构建其唯一对应的匹配规则集,如 HTTP 服务对应专门处理 HTTP 数据包的检测规则。因而,数据包检测过程中只需匹配 HTTP 检测规则集合,避免了全局遍历检测规则集合,从而大大提高了规则匹配的效率。

4 正则表达式匹配

为了能够更好地解析协议以及协议中包含的文本属性值,引入了正则表达式匹配[10]的方法用于决策树入侵检测。即将协议分析后得到的相关属性数据,采用正则表达式的方式与当前决策树分支中的属性值作匹配,从而决定最佳匹配分支。采用正则表达式匹配,

能够有助于决策,进而提高了检测效率,有效地降低入侵中的漏报率。正则表达式语法规则见表 2。

表 2　正则表达式语法规则

符号	说明
\	转义符,与字符组合表示特定的意义
*	匹配 0 次或多次
+	匹配 1 次或多次
?	匹配 0 次或 1 次
O	匹配子表达式
A/B	匹配 A 或者 B
AB	匹配 AB,其中 B 紧跟 A
[a-zA-Z]	匹配带区间的字符集
\d	匹配数字字符。等效于[0~9]
\s	匹配任何空白字符,包括空格、制表符、换页符等
\S	匹配任何非空白字符
\w	任意一个字母或数字或下划线
^,$	与符串开始(结束)的地方匹配不匹配任何字符
{}	表达式至少重复 *m* 次,最多重复 *n* 次

5　实验测试

实验环境为:CPU:Intel Core Duo 2.66GHz,操作系统:Windows XP Service Pack 2,内存 DDR2 2GB,训练和测试数据为采用 Ethereal 1.0.3 捕获的网络数据包。

实验 1 正则表达式匹配有效减少误报和漏报。

实验模拟以 HTTP 请求入侵为例,典型的攻击数据包:

"Get /winnt/system32/cmd.exe? /c + dir http/1.1" ①

实验过程:(1)基于数据包①,手工构建若干变形的攻击数据包,如"Get /winnt/system32cmd.exe? /c+dir' http/1.0"等 29 条类似的变种;(2)构造简单模式匹配的匹配规则,如 Get content:cmd.exe alert("Intrusion");(3)针对实验数据包,构造正则表达式;(4)分别应用简单模式匹配和正则表达式匹配方法对构造的 30 条攻击数据包进行测试。

实验结果:使用简单模式匹配方法仅能对攻击数据包①进行正确判断,检测其为入侵,其他的变形均为误判;而使用正则表达式匹配则对所有的测试数据包均作出了正确的入侵判断。

实验 2 基于协议分析和决策树技术的检测引擎与基于规则的简单模式匹配的性能分析。

1) 检测时间比较测试:首先通过 Ethereal 捕获足够的网络数据包,手工随机地向其中的部分数据包中添加入侵特征值,作为实验的测试数据;然后构建 C4.5 决策树生成模块,训练重组后的测试数据集,生成决策树,并根据决策树中包含的 Service 属性值生成相应的分类检测规则;最后,分别使用两种待检测算法对测试数据进行检测,记录测试结果。实验模拟过程中,具有入侵和非入侵特征的测试数据包数量相当,使得测试结果具有一定的普遍性和正确性,检测时间比较如表 3 所示。

表 3　检测时间比较　　单位:ms

测试数据量	匹配规则数	基于规则的简单模式匹配	基于协议分析和决策树	比例
30	3	0.757585	1.869228	0.405293
30	10	1.708834	1.907871	0.895676
30	30	2.832394	1.930373	1.46728
60	30	5.475570	3.295207	1.66168
60	60	10.291548	3.604008	2.85558
60	120	19.911466	3.685393	5.40281
60	240	39.067804	4.157946	9.39594
120	240	76.668177	6.672199	11.4907
240	240	151.003195	12.055587	12.5256

实验结果:在测试数据相同的前提下,当规则数量较少时,采用基于规则的简单模式匹配技术所需检测时间略少;随着规则数量的逐步增加,基于规则的简单模式匹配所需检测时间大量增加,而基于决策树的检测只出现细微的增长。当测试数据和检测规则数增加到一定程度时,基于决策树的检测比基于规则的简单模式匹配所需时间大大减少。

2) 正误报率比较测试:选取 30 条测试数据,针对性地设置 15 条检测规则;然后,手动将规则中的入侵特征值随机地添加到测试数据包的数据部分中;最后,分别使用两种待检测算法进行测试,记录测试结果。

实验结果:由于加入的入侵特征值仅在数据包的特定字段内有特殊意义,而简单模式匹配对整个数据包进行匹配,出现了大量的误判,仅对 6 条有效入侵作出了正确判断;而使用协议分析和决策树方法对所有的 30 条测试数据均进行了正确的判断。

6　结束语

采用决策树 C4.5 经典算法实现了基于决策树的入侵检测,将 Ethereal 协议分析与决策树算法有效结合,在提高检测效率以及降低误报率方面取得了较好的效果。同时引入了正则表达式匹配的思想用于协议分析和入侵检测决策树的分类匹配,不仅完善了入侵检测决策树的构建,也进一步降低了入侵检测的漏报率和误报率。

然而,基于决策树的入侵检测很难准确检测训练数据中不包含的新的入侵类别,因此,如何加强训练数据的质量以尽可能全面地包含各种入侵类型,是决策

(下转第 185 页)

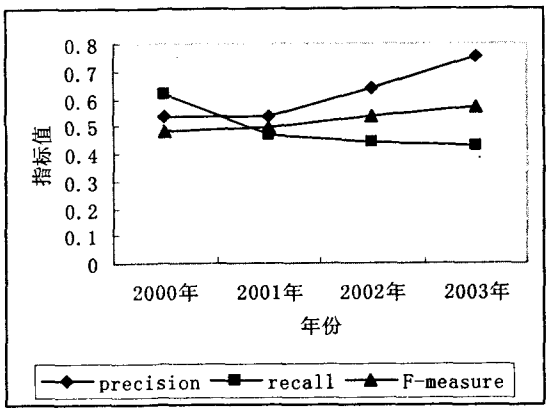

图2　系统性能曲线图

正确率和召回率在一定程度上是一对相对矛盾的指标,正确率高意味着召回率低。为了平衡两者,通常采用综合评价指标 F－measure:

$$F-measure=\frac{2*precision*recall}{precision+recall}$$

从图2可以看出,随着时间段的推移,推荐系统的正确率有一定幅度的上升,而召回率有所下降,但系统的综合评价指标 F－measure 值随着时间段的推移有所上升,这说明推荐系统的总体性能随时间段的推移有所提高。

5　结束语

文中提出的基于内容过滤的电子商务推荐系统,利用向量空间模型挖掘用户独特的兴趣需求,再根据产品信息特征的量化值来进行商品的推荐,并根据用户的反馈信息自适应学习,以提高系统的综合性能。

基于向量空间模型的方法根据特征项在文本中出现的位置和频率来判断其对文本内容的贡献,其本质只是一种浅层次的文本统计方法。基于内容过滤的推荐需要积累足够数量的评价才能构建出一个可靠的分类器,对于新用户和新产品就比较难处理;并且当用户的兴趣特征发生转变时,推荐系统模型不易更新,系统最终还要通过在实践中应用而不断优化。

参考文献:

[1] Lawrence R D, Almasi G S, Kotl Y V, et al. Personalization of supermarket product recommendations [R]. US: IBM, 2000.

[2] 吴辉娟,袁　方.个性化服务技术研究[J].计算机技术与发展,2006,16(2):32－34.

[3] 李　鹏,汪东升,陈　康.一个基于 VSM 的个性化信息推荐系统[J].计算机工程与设计,2003,24(10):19－22.

[4] 周文刚,马占欣.基于代理的 Web 页访问语义过滤与内容重现[J].计算机技术与发展,2007,17(4):120－124.

[5] 李石君,李　洲,余　军,等.基于 URL 过滤与内容过滤的网络净化模型[J].计算机技术与发展,2006,16(1):5－7.

[6] Buckley C, Salton G, Allan J, et al. Automatic query expansion using SMART[R]. [s.l.]:[s.n.],1995.

[7] 曹　毅,贺卫红.基于向量空间模型的信息安全过滤系统[J].计算机工程与设计,2006,27(2):224－227.

[8] 黄萱菁,夏迎炬,吴立德.基于向量空间模型的文本过滤系统[J].软件学报,2003,14(3):435－442.

[9] Robertson S, Hull D. The TREC－9 filtering track final report [C]//In: Voorhees E M, Harman D K, eds. Proceedings of the 9th Text Retrieval Conference (TREC－9). Gaithersburg: NIST Special Publication, 2001:25－40.

(上接第181页)

树检测效率的一个重要因素。通过及时对训练数据进行补充更新,以便将新的入侵特征更新到决策树模型中。

参考文献:

[1] 凌　宇,徐　雄,石林安. NIDS 中协议分析和模式匹配的研究[J]. 信息技术, 2006(5):82－84.

[2] 蔡　敏,叶　震,徐吉斌. 协议分析技术在入侵检测中的应用[J]. 计算机技术与发展,2007,17(2):240－241.

[3] Lee W. A data mining framework for constructing features and models for intrusion detection systems[D]. New York: Columbia University, 1999:250－310.

[4] 李庆华,赵延喜,蒋盛益. 基于数据挖掘的协议分析检测模型[J]. 计算机工程与设计,2005,26(7):1701－1703.

[5] 王丽萍,孙　蕾. 基于 Ethereal 开源代码构建协议解析器的方法研究[J]. 计算机技术与发展,2007,17(10):27－30.

[6] 杨学兵,张　俊. 决策树算法及核心技术[J]. 计算机技术与发展, 2007,17(1):43－45.

[7] Han Jian－wei, Kamber M. Data mining concepts and techniques[M]. Beijing: China Machine Press, 2000:180－200.

[8] Lee W, Stolfo S J, Mok K W. Data mining in workflow environments: experiences in intrusion detection [C] // ACM SIGKDD International Conference on Knowledge Discovery & Data Mining(KDD－99). USA:[s.n.],1999:114－124.

[9] Kruegel C, Toth T. Using decision trees to improve signature based intrusion detection[C]//Proc of the 6th International Workshop on the Recent Advances in Intrusion Detection (RAID). USA: Springer－Verlag, 2003:173－191.

[10] 佘石泉,周肆清. 正则表达式在编程题自动阅卷中的应用[J]. 计算机技术与发展, 2007,17(7):244－246.