

文章编号:1001-9081(2009)02-0409-03

## 基于数据挖掘的 Snort 系统改进模型

张亚玲,康立锦

(西安理工大学 计算机科学与工程学院, 西安 710048)

(ylzhang@xaut.edu.cn; kanglijin2006@163.com)

**摘要:**针对 Snort 系统对新的入侵行为无能为力的缺点,设计了一种基于数据挖掘理论的 Snort 网络入侵检测系统的改进模型。该模型在 Snort 入侵检测系统的基础上增加了正常行为模式挖掘模块、异常检测引擎模块和新规则生成模块,使得系统具有从新的入侵行为中学习新规则和从正常数据中学习正常行为模式的双重能力。实验结果表明,新模型不仅能够有效地检测到新的入侵行为,降低了 Snort 系统的漏报率,而且提高了系统的检测效率。

**关键词:**入侵检测; Snort 系统; 数据挖掘; 规则学习

**中图分类号:** TP393.08 **文献标志码:** A

## An improved model of Snort system based on data mining

ZHANG Ya-ling, KANG Li-jin

(School of Computer Science and Engineering, Xi'an University of Technology, Xi'an Shaanxi 710048, China)

**Abstract:** An improved model of the Snort network intrusion detection system based on the theory of data mining was proposed, regarding the problem that Snort is powerless to new types of intrusion. In the new model, normal behavior patterns mining module, anomaly detection engine module and new rules generating module were added to the Snort system. By these improvements the system has double capacity of learning rules from new intrusions and learning normal behavior patterns from normal data. The test result shows that new types of intrusion can be detected effectively, the false negative of Snort can be reduced, and the detection efficiency of the system has been enhanced.

**Key words:** intrusion detection; Snort; data mining; rule learning

### 0 引言

入侵检测技术通过从计算机网络或计算机系统中的若干关键点收集信息并对其进行分析,从中发现网络或系统中非授权的或者威胁到系统安全的行为,同时对该行为做出响应,达到保证系统安全的目的。根据检测方法的不同,入侵检测可以分为误用检测和异常检测两种。目前大多数网络入侵检测系统都是事先设定好入侵规则库,然后通过规则匹配来发现入侵数据,也就是所谓的误用检测机制。这一做法的最大缺点就是不能发现新的入侵行为,漏报率较高。最受关注和广泛使用的开源入侵检测系统——Snort 就是一个典型的误用型入侵检测系统,其最大缺点也就不言而喻。

数据挖掘(Data Mining, DM)是从大量数据中提取或“挖掘”知识,它能高度自动化地分析原有的数据,做出归纳性的推理,从中挖掘出潜在的模式,预测出对象的行为。数据挖掘的定义有广义和狭义之分。从广义的观点,数据挖掘是从大型数据集(可能是不完全的、有噪声的、不确定性的、各种存储形式的)中,挖掘隐含在其中的、人们事先不知道的、对决策有用的知识的过程。从狭义的观点,可以定义数据挖掘是从特定形式的数据集中提炼知识的过程<sup>[2]</sup>。

Wenke Lee 于 1999 年首次将数据挖掘技术引入入侵检测,它把入侵检测看成是一个数据分析过程,利用数据挖掘技术自动构建系统的特征模式,提高了入侵检测系统的准确性、扩展性和自适应性,成为入侵检测研究领域的一个新的研

究热点,国内外许多学者都进行了大量的研究。目前,数据挖掘中的关联分析算法使用最多,它可以被用来挖掘关联模式,进行异常检测<sup>[3]</sup>。序列模式挖掘算法可以被用来发现数据的前后关系,挖掘出序列模式<sup>[4]</sup>。聚类算法也可以被用来构建正常行为模式<sup>[5]</sup>,进行异常检测。还有分类算法可以用来构造分类器<sup>[6]</sup>,分类器经过大量的入侵数据集训练之后可以用于入侵检测。本文以提高 Snort 的检测效率和增强 Snort 系统对新入侵行为的检测能力为目标,运用数据挖掘理论将误用检测和异常检测相结合,设计并实现了一种基于数据挖掘的 Snort 系统的改进模型。

### 1 Snort 系统的入侵检测模型分析

Snort 是一个开源网络入侵检测系统,由 Sourcefire 公司 Martin Roesch 等人开发,当前最新版本是 Snort 2.8。近几年 Snort 取得了快速的进展,是目前安全领域最活跃的开源网络入侵检测系统之一。Snort 提供的插件机制使得其扩展起来非常方便,因此许多企业在 Snort 的基础上进行二次开发,扩展 Snort 的功能,以适应不同的网络安全需求。同时,Snort 简洁的结构、高效的代码使其成为众多学者进行入侵检测学习和研究的对象。

Snort 可以分为 5 个主要的组件,每个组件对入侵检测都很关键。第一个是捕包装置,Snort 依赖一个外部捕包程序库 libpcap 来抓包。在包以原始状态被捕获后,要送给包解码器。解码器是进入 Snort 自身体系的第一步,包解码器将特殊

收稿日期:2008-08-06;修回日期:2008-09-23。

基金项目:教育部科学技术研究重点项目(208139);陕西省自然科学基金资助项目(2006F37)。

作者简介:张亚玲(1966-),女,陕西西安人,副教授,主要研究方向:密码理论、网络安全;康立锦(1982-),女,陕西咸阳人,硕士研究生,主要研究方向:网络与信息安全。

协议元素翻译成内部数据结构。在最初的捕包和解码完成后,由预处理程序处理流量。许多插入式预处理程序对包进行检查或操作后将它们交给下一个组件:检测引擎。检测引擎利用事先设置的规则库对每个包进行检测来发现入侵。最后一个组件是输出插件,它对检测出来的入侵行为产生报警。其系统结构如图1所示。

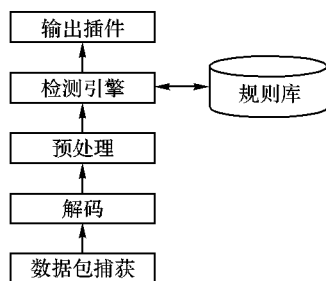


图1 Snort系统结构

从图1中可以看出,Snort系统是通过一个已有的规则库进行入侵行为的检测,其中没有规则的扩充机制,这就使得它对于新的攻击行为无能为力;此外,根据美国国防部高级研究规划局(DARPA)数据显示,网络流量中95%以上的数据是正常数据。也就是说,网络上正常的数据包占有非常大的比例,而攻击数据包的数量是有限的,因此可以考虑在进入检测引擎之前先排除大部分正常数据包,以降低检测引擎的负担,达到提高效率的目的。

## 2 基于数据挖掘的Snort系统改进的模型

Snort系统属于误用型网络入侵检测系统,其检测能力依赖于规则库。如果规则库得不到及时快速的更新,那么Snort将不能检测到新的入侵行为,这样必然导致Snort系统漏报率增加。针对这一问题,本文利用数据挖掘技术将异常检测技术融入到Snort原有系统,设计了一种基于数据挖掘的Snort系统的改进模型,以提高Snort的检测效率并降低其漏报率。

具体设计思想:首先,利用数据挖掘技术构建正常行为模式库,并在Snort检测引擎之前加入异常检测引擎来过滤掉大部分正常数据,减少Snort检测引擎的负担,提高其检测效率;然后,对可疑数据作进一步处理,使新的入侵数据分离出来并利用数据挖掘算法对其进行关联分析,最终生成适合Snort检测的新规则,存入Snort规则库;这里的可疑数据是指过滤掉大部分正常数据和已知入侵数据之后的数据(可能包括部分正常数据和新的入侵数据)。最后,将分离出来的正常数据存入正常数据库,以便不断更新正常行为模式库。这一模型可以起到两方面的作用:1)提高Snort的检测效率,因为大量的正常数据包被提前过滤;2)能够发现并检测到新的入侵行为,因为Snort规则库可以得到不断的更新。

根据上述设计思想,本文完成的基于数据挖掘的Snort系统改进模型如图2所示。

- 1) 正常行为模式挖掘模块:主要利用数据挖掘算法来挖掘出正常行为模式,构建正常行为模式库,为异常检测引擎模块做预备工作。
- 2) 异常检测引擎模块:利用前面产生的正常行为模式来检测异常数据,快速过滤掉大部分正常数据以减轻Snort检测引擎的负担。
- 3) 生成新规则模块:对可疑数据进行分类,并利用关联分析算法对分离出来的新入侵数据进行关联分析,最终生成适合Snort检测的新规则。

图3给出了改进模型的系统数据流。可以看到,异常检测引擎利用经过挖掘过程得到的正常行为模式库可以快速过滤掉大量正常数据;经过分类和新规则生成模块得到的新入侵规则可以不断更新Snort规则库,使Snort系统能够检测到新的入侵行为。通过两方面的改进,Snort系统的检测效率不仅得到了提高,而且能够不断扩充Snort规则库,有效地检测到新的入侵行为。

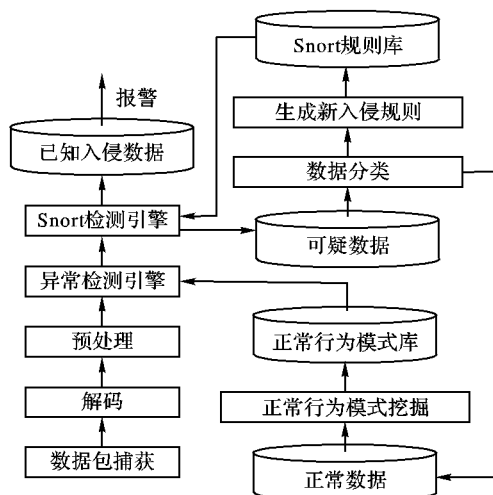


图2 基于数据挖掘的Snort系统的改进模型

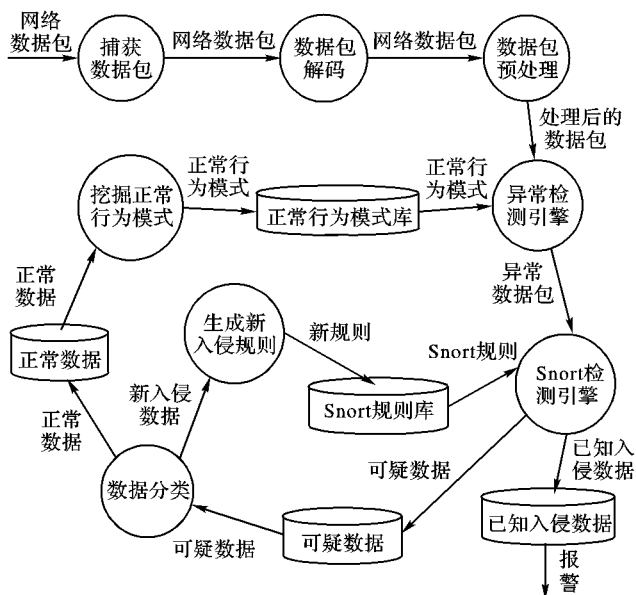


图3 基于数据挖掘的Snort系统的改进模型数据流

## 3 系统实现

本文改进模型的实现是基于开源网络入侵检测系统Snort及其相关组件的,这些都可以通过互联网免费获得。具体实现过程分为以下几步进行:

- 1) 在Windows系统下部署一个简单的Snort网络入侵检测系统。用到的主要软件有:Snort、Winpcap、jgraph、Mysql、Apache、php、ADODB、ACID等,对这些软件进行合理的安装和配置,构建起一个功能较完善的Snort入侵检测系统。
- 2) 将可疑数据存入可疑数据库并对其进行分类,使新的入侵数据和正常数据分开。这部分通过在ACID(Analysis Console for Intrusion Databases)中加入相应的处理页面来人机交互实现。
- 3) 利用数据挖掘算法构建正常行为模式库并实现异常

检测引擎模块。本文对数据挖掘算法中的关联规则算法 Apriori 算法进行适当改进,使其挖掘出正常行为模式。异常检测引擎模块在 Snort 程序中使用 C 语言来实现。

4) 利用关联规则算法 Apriori 算法对分离出来的新入侵数据进行关联分析,生成关联规则并转换成适合 Snort 检测的新规则。这部分在 ACID 中使用 PHP 语言来实现。

## 4 模拟实验及结果

为了对该模型的检测效果进行评估,本文进行了一些相关的模拟实验。本实验分为两个部分,其一是对于改进效率的测试,其二是对于规则扩充能力的测试。

实验方法:利用 Profile 工具来进行实验。VC++ 6.0 提供的 Profile 工具可以帮助程序员发现程序运行的瓶颈,找到耗时所在。

实验数据:由于在局域网环境下,每一时间段的网络流量不同,来自局域网的实际网络数据不能准确评价系统改进前后的效率,因此,本文采用美国麻省理工林肯实验室 (MIT Lincoln Lab) 提供的 1999 DARPA Intrusion Detection Evaluation Data Sets<sup>[7]</sup> 来进行对比实验,以说明改进效果。

用网络分析软件 Ethereal 打开数据集 outside.tcpdump,共有 233 428 条记录,本文就该数据集的前 1 000、5 000 和 10 000 条记录对改进前后的系统进行实验,给出改进前后系统的总耗时和\_DETECT 函数调用次数的对比,结果如表 1 所示。

表 1 改进前后检测效率对照表

数据记录/条	改进后总耗时/ms	改进前总耗时/ms	改进后调用_DETECT函数次数	改进前调用_DETECT函数次数
1 000	24 085.952	24 265.795	788	986
5 000	94 590.281	95 201.430	3 442	5 138
10 000	326 387.279	329 169.852	5 016	10 409

从上面的实验结果可以看出,改进后的系统总耗时明显减少,而且调用 Snort 原有检测函数\_DETECT 的次数也明显减少。这就说明改进后的系统减轻了 Snort 原有检测引擎的负担,提高了 Snort 系统的检测效率。

对新规则生成模块进行实验的过程中发现,网络数据的特定形式不适应 Apriori 算法直接进行挖掘。因为,源 IP 和目的 IP 属于两个不同的项,但值可能相同,源端口和目的端口也存在这样的问题,直接对这些数据进行挖掘,将产生大量冗余的频繁项集和强关联规则。因此,必须对数据进行预先处理,本文的做法是对每条记录的源 IP 和源端口加入标识符 s,目的 IP 和目的端口加上标识符 d。然后,利用关联分析算法 Apriori 算法对处理后的数据进行挖掘,设定最小支持度和最小置信度分别为 10% 和 98%,改进系统在对 50 条新入侵数据进行挖掘之后,自动生成了 3 条规则并存入 Snort 规则库 rules 下的 new.rules 中,如下所示:

```
alert TCP 172.16.117.52 1051 -> 207.46.130.139 80
```

```
alert TCP 172.16.117.52 1049 -> 207.46.130.139 80
```

最后,重新运行系统,发现 Snort 在初始化时已经将这三条规则加入到规则链中,并且对测试数据中匹配到该规则的数据进行报警,相应的报警信息可以在报警文件 alert.ids 中找到,而且在 ACID 分析控制台也可以查看到这些报警信息,图 4 中给出了其中的部分报警信息:

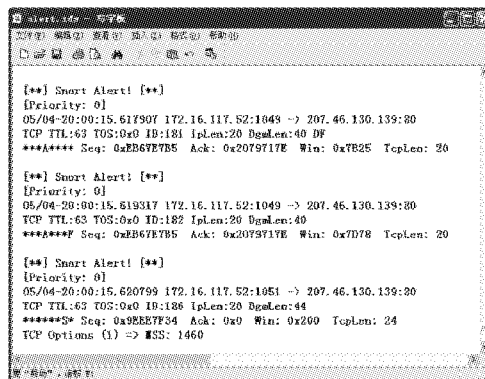


图 4 新规则生效后的报警信息

上面的测试表明:该系统可以生成新的入侵规则并扩充 Snort 的规则库,因而改进后的 Snort 系统对新类型的入侵行为为具有了一定的防御能力。

## 5 结语

本文基于数据挖掘理论,提出了一种 Snort 系统的改进模型,并通过 Snort 的插件机制实现了该模型。经过采用美国麻省理工林肯实验室 (MIT Lincoln Lab) 提供的测试数据集进行改进前后的对比实验,结果表明该模型不仅提高了 Snort 系统的检测效率,并且可以不断扩充 Snort 的规则库,使其能够检测到新的入侵行为。本文对数据挖掘技术在入侵检测系统中的应用进行了初步尝试,更进一步的研究工作如挖掘项对于检测效果的影响、最小支持度和最小置信度对模式库精确度的影响等还有待深入。

### 参考文献:

- [1] LEE W, STOLFO S J. A framework for constructing features and models for intrusion detection systems[J]. ACM Transactions on Information and System Security, 2000, 3(4): 227 - 261.
- [2] HAN JIAWEI, KAMBER M. 数据挖掘概念与技术[M]. 范明, 孟小峰, 译. 北京: 机械工业出版社, 2001.
- [3] 陈耿, 朱玉全, 孙志辉, 等. 一种基于异常检测的关联模式挖掘模型[J]. 计算机工程与应用, 2004, 40(12): 158 - 198.
- [4] 宋世杰, 胡华平, 胡笑蕾, 等. 数据挖掘技术在网络型异常入侵检测系统中的应用[J]. 计算机应用, 2003, 23(12): 20 - 23.
- [5] 贾世国, 张昌城等. 基于数据挖掘的网络入侵检测系统的设计与实现[J]. 计算机工程与应用, 2008, 44(14): 134 - 137.
- [6] 宋世杰, 胡华平, 胡笑蕾, 等. 数据挖掘技术在网络型误用入侵检测系统中的应用[J]. 计算机工程, 2004, 30(16): 126 - 127.
- [7] MIT Lincoln Lab. Intrusion detection evaluation data sets. [EB/OL]. [2008 - 06 - 10]. <http://www.ll.mit.edu/mission/communications/ist/corpora/ideval/data/1999data.html>.

(上接第 408 页)

- [9] 杨扬, 孙志伟, 赵政. 一种处理障碍约束的基于密度的空间聚类算法[J]. 计算机应用, 2007, 27(7): 1688 - 1692.
- [10] 薛丽霞, 汪林林, 王佐成, 等. 基于 Voronoi 图的有障碍物空间聚类[J]. 计算机科学, 2007, 34(2): 189 - 192.
- [11] 严馨, 周丽华, 陈克平. 一种改进的带障碍的基于密度和网格的聚类算法[J]. 计算机应用, 2005, 25(8): 1818 - 1822.
- [12] ZHANG XUEPING, WANG JIAYAO, FAN ZHONG SHAN, et al.

Spatial clustering with obstacles constraints using ant colony and particle swarm optimization [C]// Proceeding of PAKDD 2007 Workshops. Berlin: Springer, 2007: 344 - 356.

- [13] GRYGORASH O, ZHOU YAN, JORGENSEN Z. Minimum spanning tree based clustering algorithms [C]// Proceedings of 18th IEEE International Conference on ICTAI. Washington: IEEE Computer Society, 2006: 73 - 81.