基于特征的入侵检测系统的评估新方法[1]

孙美凤[1,2]　龚俭[1]　　杨望[1]

[1]（东南大学计算机科学与工程系，江苏省计算机网络技术重点实验室，江苏 南京 210096)

[2]（扬州大学信息工程学院计算机系，江苏 扬州 225000)

A new approach to evaluate the capacity of signature-based intrusion detection systems

Sun Meifeng[1,2]　Gong Jian[1]　Yang Wang[1]

[1] (Department of Computer Science and Engineering, Southeast University, Nanjing 210096, China)

[2] (College of Information Engineering , Yangzhou University, Yangzhou 225000, China)

Abstract: Existing intrusion detection system (IDS) evaluation methods take an IDS as a black-box, and deduce its detection capabilities by observing its outputs under a traffic mixture of normal usages and attacks. The results by such an evaluation method reflect the capacity of a signature-based IDS, which is determined by its implementation combined with human knowledge input in it. Since the detection rule format and its semantic definition may vary, the precondition for the evaluation is not equal in fact. Therefore, the current methods are not reasonable enough, and the results may change as the detection rule changes. In this paper, we propose a new evaluation method for signature-based IDS, which views the human knowledge as IDS parameters, and evaluates the capability of IDS implementation only. We focus on the definition of metrics. Additionally, we also introduce how to calculate the value of metrics. A prototype is implemented which shows that this new method can evaluate the real capacity better for a signature-based IDS.

Keywords: intrusion detection, signature-based intrusion detection system, evaluation

摘　要：目前评估方法将 IDS 看作一个黑盒子，通过观察它在模拟的正常用户行为和入侵作用下的输出来推断其在实际环境中**表现出的检测能力**。对基于特征的 IDS，这种表现出的检测能力反映 IDS 实现和预置人工知识的综合质量。由于 IDS 各自定义所使用的检测规则，并且在定义之后，规则及其数量还可能变化，所以 IDS 在评估时表现出的检测能力与实际运行中表现出的能力可能不同，这就失去了评估的意义。本文提出一种基于**系统能力**的评估方法，该方法把人工知识视为评估参数，因此结论反映 IDS 实现的质量。本文重点讨论系统能力的测度定义，并简单介绍测度计算的总体思路。实验结果表明本文方法更能反映基于特征的 IDS 的真实质量。

关键字：入侵检测；基于特征的入侵检测系统；评估

中图分类号：TP393

1　引言

20世纪80年代以来，随着对入侵检测技术研究的发展，出现了许多入侵检测系统（IDS），

[1] 本文受国家 973 计划课题（2003CB314804）、教育部科学技术重点研究项目(105084)和江苏省网络与信息安全重点实验室（BM2003201）资助。作者简介：孙美凤(1970-)，女，博士生，研究方向为入侵检测系统，EMAIL：msun@njnet.edu.cn；龚俭(1957-)，教授，博士生导师，研究方向包括网络安全，网络行为学；杨望（1979－），男，博士生，研究方向位入侵检测系统

因此产生了对各种入侵检测系统功能和性能评测的需求。无论采用何种检测方法，IDS 的最终目的都是应用到某一实际环境中进行入侵检测。具体说，就是从实际环境输入的反映入侵和正常行为的审计数据流中识别出入侵，然后报警。因此 IDS 用户，特别是潜在客户，期望了解 IDS 的实际检测能力（包括识别入侵和通过正常行为的能力，通常用漏报率和误报率描述）。从这个角度，现有评估工作都以**表现出的检测能力**为评估目标。为了评估的可控性、可重复性和攻击标定等要求，评估不可能在真实环境中进行。因此现有研究工作都围绕着真实环境的模拟，并且其成功与否依赖测试数据是否准确刻画了真实环境。

这种评估表现出的检测能力的观点始见于加州大学戴维斯分校安全实验室（以下简称 UCD）1995 年的技术报告[1]，由于麻省理工学院林肯实验室（以下简称 MIT LL）[2~4]在 1998～2000 之间的工作而开始引人关注。后续的评估工作[5,6]以及工业界的各种产品评估都基于该观点进行[7~9]。综述文献[10~12]对其中主要工作有详细介绍。

McHugh 对 MIT LL 的工作提出尖锐批评[13]，主要诟病是：（1）按攻击者意图进行的 Probe/u2r/r2l/Dos 分类与检测方法无关，因此基于该分类的入侵选择缺乏恰当的理论基础；（2）背景流量的建模方法主要依据专家经验和粗略的统计模型，不能准确地刻画实际网络环境，尤其实际网络环境的特征随着不同的地理区域、用户群和应用程序表现出很大的差异。总之，McHugh 认为 MIT LL 采用的技术粗糙，不能保证测试数据的代表性，评估结论依赖测试环境。

实际环境中的入侵和正常行为表现出多维的属性。与属性多样性对应，IDS 检测机理也各不相同，即表现出“聚焦”某一入侵属性而“忽略”其它属性的特点。为了公平，测试数据应该覆盖实际环境的各种属性，否则可造成对某种检测机理的偏向，从而使评估结论依赖测试数据。然而入侵和正常行为是客观现象，从知识表示的角度，除了客观对象自身，与客观对象完全一致的表示在理论上是不存在的[14]。因此如果不引入检测机理的知识，单纯地依靠模拟技术试图得到一个对 IDS 广泛适用的测试数据集，即使有可能也必然非常困难。所以，本文认为 MIT LL 测试数据集的根本问题不是技术粗糙，而是它设定了一个不现实的研究域。

因此，我们可能不应该期望一个对所有 IDS 通用的测试方法和测试数据集，而应该探讨如何对 IDS 的检测机理进行分类，并应用检测机理的知识简化问题。例如 IBM 研究过异常检测系统的评测工作[15]，所提出的 3 种模拟背景流量的方法被公认具有较高的学术价值。本文则面向基于特征的 IDS（包括滥用检测系统和目前热点中的关联分析系统）。尽管基于特征的 IDS 有维护工作量大并且不能识别新入侵的缺点，但是它的误报率低，并且报警结论可为管理员提供直接的指导，因此在生产实际中得到广泛应用，研究对它的准确评估具有实际价值。

基于特征的 IDS 所表现出的检测能力与规则库内容（入侵特征）有关。由于不同的 IDS 各自定义规则库，使得这些规则库内容可能不同，因此检测能力可能是在不平等条件下表现出来的。如果各 IDS 的规则库内容固定不变，那么规则库内容就成为 IDS 实现的一部分，直接评估表现出的检测能力未尝不可。然而所有 IDS 都允许管理员添加、删除和修改规则，因此各 IDS 评估时的能力在实际使用过程中可能发生变化，导致评估结论仅在评测时有效，这就失去了评估的意义。从这个角度，评估表现出的检测能力的观点，即使存在完善的技术，也不适合基于特征的 IDS，而以前的研究中从未提到这一点。本文则从基于特征的 IDS 的检测原理出发，提出一种评估新方法，它评估 IDS 的**系统能力**。这种系统能力表现为 IDS 系统功能及其实现质量决定的对入侵特征的处理能力，是 IDS 功能和性能的稳定特征。在这个概念下，入侵特征被视为评估参数。当给予相同的入侵特征，系统能力强的 IDS 一定表现出更好的实际检测效果；反之，即使有完善的入侵特征，系统能力弱的 IDS 也可能表现出不好的检测效果，例如不能准确地解释规则，或性能跟不上。因此系统能力综合反映 IDS

表现出的和潜在的检测能力。

本文结构如下：首先给出本文方法的评估对象即系统能力的定义；然后在第 3 节讨论测度的选取；第 4 节给出测度值计算的总体思路；作为示例，第 5 节对开源系统 Snort、Bro 以及作者所在实验室开发的一个入侵检测系统 Monster 进行测试并给出测试结果；最后是本文的结论。

2 系统能力的定义

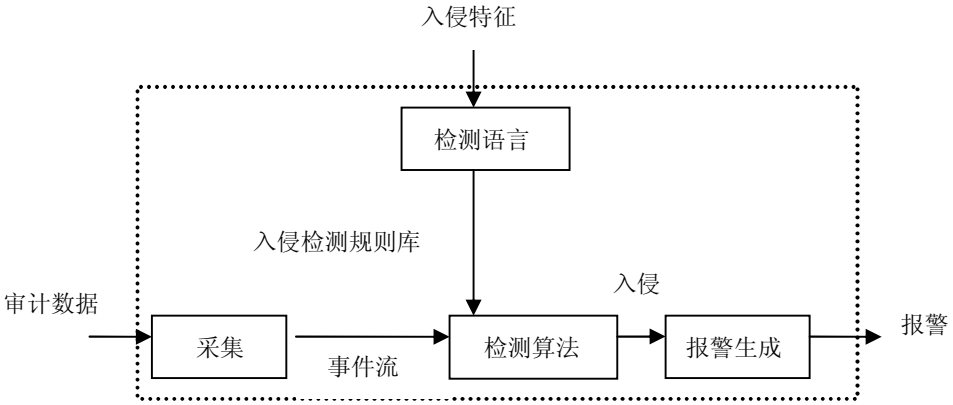

图 1 基于特征的 IDS 的功能模型

基于特征的 IDS 假定入侵具有可区分的特定行为模式，该行为模式在系统中的痕迹模式称为入侵特征。把一个用户行为的痕迹称为事件，则入侵特征形式上表现为事件约束和事件间关系的约束。审计数据是事件的物理存储，通常包括主机日志和网络报文，前者是 HIDS（基于主机的 IDS）的数据源，后者是 NIDS（基于网络的 IDS）的数据源。如同高级语言程序必须被翻译成机器语言程序才能执行一样，抽象的入侵特征必须被表示成检测语言的规范形式才能用于检测，这项功能通常由 IDS 管理员手工完成。运行中的 IDS 不断从审计数据中采集事件，检测算法依据入侵检测规则在事件流中寻找恶意行为的证据，并在发现入侵证据时触发报警。因此，对基于特征的 IDS，表现出的检测能力由 2 个相互独立的要素决定：**入侵特征的描述能力**和 **IDS 的系统能力**。

入侵特征的描述能力指入侵特征相对于它所描述的入侵的完备性和精确性（对应漏报和误报），从不同的角度观察同一个入侵可以得到完全不同的特征，它们的描述能力也各不相同。入侵特征的抽取是一个专门的研究热点，但不在本文的讨论范围。一个 IDS 配置什么样的入侵特征是由管理员决定的，并且在配置之后，管理员还可以通过增加、删除和修改的操作不断调整 IDS 和入侵特征的绑定。这种定义和使用的方式说明入侵特征是独立于 IDS 的因素，是 IDS 的环境之一。

如果将入侵特征看作是向 IDS 下达的“入侵检测命令”，则图 1 虚线框内的 4 个子功能相互合作，负责“命令”的执行。出于资源和效率的考虑，这些子功能可能一定程度地牺牲正确性，导致 IDS 表现出的检测能力与理想能力之间的误差。例如：否定语义对提高入侵特征的精确性很有用，但多数检测语言不支持它，从而抽象入侵特征向入侵检测规则变换的过程可能丢失语义。由于 IDS 各自定义和实现每个子功能，并且这些子功能在软件实现后通常不能改变（软件升级后的 IDS 已经不是原来的 IDS 了），因此图 1 中虚线框内的功能实体是 IDS 固有的，是 IDS 真正的组成部分，它们综合表现出的“命令”执行能力就是 IDS 的系统能力。

定义 1 IDS 的系统能力指由检测语言的表达能力、事件采集能力、检测算法的分析能

力以及报警能力共同决定的 IDS 对入侵特征执行的准确性 $\sum_{A,S}|f_{IDS}(A,S)-f_{理想}(A,S)|$，其中，$A=(a_1,a_2,...,a_n)$ 是某时间区间的用户行为序列，S 是入侵特征集，$f_{IDS}(A,S)$ 和 $f_{理想}(A,S)$ 分别指实际 IDS 和理想 IDS 作用于 A 和 S 得到的入侵事件。

为了有效执行入侵特征，IDS 首先必须正确“理解”入侵特征的内容，即在将抽象的入侵特征翻译成检测语言规范格式（入侵检测规则）的过程中不丢失信息，这由检测语言的表达能力决定。其次，IDS 必须拥有“好视力”，即采集到的事件流包含同一时间实际发生的用户行为的全部细节，事件采集能力决定两者的误差大小。检测算法的分析能力是检测算法的实现与理想检测算法之间的误差。报警能力对应于检测算法实现捕捉到的入侵证据和最终报警数据的误差。

评估 IDS 应以系统能力为对象，因为它抓住了 IDS 实现的本质，综合反映出 IDS 表现出的和潜在的检测能力。“好系统＋好特征”的组合一定表现出好的检测能力；“坏系统＋坏特征”的组合一定表现出坏的检测能力。由于好特征的威力或者好系统的能力发挥不出来，“坏系统＋好特征”和“好系统和坏特征”的组合可能都表现出坏的实际检测能力，但前者表现出的坏能力是 IDS 固有的，只能通过升级 IDS 解决；而后者表现出的坏能力则是临时性的，传统方法无法揭示这两种组合的本质不同。

3 系统能力的测度定义

按照第 2 节，系统能力由 4 个子能力组成：检测语言的表达能力、事件采集能力、检测算法的分析能力和报警能力。但是，从评估的角度，如果 IDS 表现出对某类入侵特征缺乏支持，我们无法严格区分是检测语言的表达能力不够或是检测算法对入侵检测规则使用时的简化造成。因此测度与子能力并不一一对应，而是全部能力的外在表现。

最后，本文不规定任何特定的实现技术，讨论的基础是图 1 中的 IDS 抽象模型。

3.1 功能测度

功能性测度刻画理想条件下 IDS 的系统能力，理想的含义是：IDS 没有受到攻击、输入规模不超过 IDS 的负荷，因此功能性测度反映 IDS 代码的有效性，即 IDS 代码是否对相关的问题给予了足够重视。

定义 2　入侵特征分辨率指 IDS 适应的入侵特征的范围

操作分辨率×时间分辨率×空间分辨率 。

由于入侵是发生在特定时间和空间的行为序列，每个行为都包括：操作、时间和空间（行为的主体和客体）属性，因此入侵特征（入侵过程的不变量）可能分布在操作、时间和空间维。直观上，IDS 需要专门的代码处理各个维的特征，否则执行时就不得不忽略该维的语义，引入执行误差。同样，维特征可能包含不同信息类型和信息量，因此面向某个维的代码功能也有不同。因此入侵特征分辨率＝操作分辨率×时间分辨率×空间分辨率，其中维分辨率代表 IDS 适应的该维特征的范围。

根据 IDS 要考虑的问题复杂程度，可以对维分辨率区分等级并量化，令较大的数值对应较复杂的问题，因而也对应了较大的维特征范围，这样入侵特征分辨率就具有了熟悉的量化形式。表 1～3 是从 Kumar 特征[16]中识别出的各维分辨率等级。Kumar 在 1995 年对攻击特征的研究工作得到了广泛认可，他基于当时主要的 UNIX 攻击，归纳出了入侵特征的结构。尽管随着互联网规模的扩大，入侵手段较 1995 年有了较大进步，基于网络的大规模、分布式协同入侵成为主要形式。但是，不管是本地攻击还是远程攻击，它们都是入侵者利用系统

的漏洞设计出来的，是入侵者思维过程的表现，工作原理和行为模式不会有根本变化。因此本文认为 Kumar 的研究在今天仍有指导意义。

显然，IDS 适应的维越多，每个维的分辨率越大，IDS 就支持对入侵过程更精确的描述，具有更强的潜在检测能力。该测度衡量检测语言的表达能力和检测算法对入侵检测规则使用的正确性。

表 1 操作分辨率

操作分辨率	说明
1	支持对入侵过程的单个正向事件（必须出现的事件）的描述
2	支持对入侵过程的多个正向事件的描述
3	支持入侵过程的正向和反向事件（不可以出现的事件）的描述

表 2　时间分辨率

时间分辨率	说明
0	不支持时间维的特征
1	支持多事件的顺序约束
2	支持多事件的时间区间长度约束

表 3　空间分辨率

空间分辨率	说明
0	不支持空间维的特征
1	支持多事件的内容关联约束

定义 3　用户行为分辨率指 IDS 采集到的事件数据准确表示用户行为痕迹的可能性

$$f(\text{事件数据的层次，理想事件的层次})。$$

在网络环境中，用户行为是应用层的概念，安全专家也在该层次上定义入侵特征，因此理想事件是应用层协议数据单元，应用层协议数据单元的序列不仅包含用户行为的内容，而且反映出行为的上下文关系。按照 TCP/IP 工作机制，TCP 把应用层协议数据单元的序列当作八位组或字节流的序列，而为了传输，又把这个序列划分成若干段，通常每个段被封装到一个 IP 报文中。显然，当取 TCP 流为事件时，IDS 准确捕捉到了用户行为的内容细节，但丢失了行为的边界或者说丢失了行为的上下文。进一步，由于 TCP 对流的划分以及中间路由器对 IP 报文的分片功能都是不考虑语义的，可能出现单个用户行为的内容被分布到多个 IP 报文的情形。这样，当取 IP 报文为事件时，IDS 不仅无法捕捉用户行为的上下文而且可能丢失内容。传统的轻量级 IDS 就直接取 IP 报文为事件，因为观测不准确造成利用 IP 碎片进行入侵逃逸的问题。如果 IDS 在采集的同时进行报文重组，在 TCP 流的层次上进行检测，就能够阻止入侵逃逸。理想情形是 IDS 具有应用层协议解析功能，能够从 IP 报文序列中重构用户行为的序列。因此 IDS 的用户行为分辨率与它视为事件的数据所在的概念层有关。推广到一般情形，IDS 的用户行为分辨率依赖于实际事件与理想事件的相对距离。

定义 4　报警信息量：报警中携带的入侵证据信息量 $\sum_i k_i d_i$，其中 $d_i \in \{0,1\}$ 表示报警中是否包含 i 维的信息，k_i 是 i 维的权值。

一次入侵包括入侵者、入侵技术、入侵目标和入侵结果等多维的信息，显然报警中提供的信息越丰富，管理员越容易准确响应，但是这些信息的生成、存储和传递需要代码的支持。该测度度量报警对响应和修复的支持程度。

3.2 性能测度

如图 1 所示，实时 IDS 对每个到达数据执行采集、分析和报警的过程，该过程的基本操作数是 IDS 代码的固有属性——时间复杂性。这意味着在一定的资源条件下，IDS 可承受的数据到达率存在极限，该极限称为 IDS 的负荷能力，它是 IDS 时间性能的外在表现。超出负荷能力的数据将被丢弃，如果丢弃数据中包含入侵证据信息，IDS 就会漏报或误报，因此 IDS 的负荷能力是可测的。

定义 5　处理效率指 IDS 的负荷能力/资源量

$$f(\text{入侵特征规模，并发入侵数，攻击密度})。$$

正如算法的时间复杂性形式上为输入大小的函数，作为 IDS 时间性能外在表现的 IDS 处理效率也可表为函数形式，通常将该函数的参数称为压力条件。由于各 IDS 的检测方法不统一，准确全面地定义压力条件是困难的。本文基于图 1 的抽象检测模型，定义可能的压力条件包括：

- **入侵特征规模**指安全专家提供的入侵知识量。对每个到达事件，检测算法在规则库中搜索匹配的规则，直观上搜索空间越大，需要的搜索时间越多。
- **并发入侵数**指同一时间段内已经发起但还没有结束的入侵数量。该条件面向内容关联从而要求穷尽搜索的多事件入侵特征，穷尽搜索要求保留特征的所有部分匹配，随着时间的推移，实际的搜索空间将大大增加。
- **攻击密度**指测试样本中，入侵数据占总样本的比值。通常入侵事件必须经过多个条件的测试，并且入侵事件要求报警处理，而非入侵事件可能在分析的早期就因为某个条件不满足被抛弃，即入侵事件和非入侵事件的处理工作量不同，从而 IDS 的负荷能力一定会受到攻击密度的影响。当测试数据为纯正常行为数据流，IDS 的负荷能力主要反映事件采集模块的性能；随着攻击密度的增加，IDS 的负荷能力将更多地受到检测和报警模块性能的影响。

影响算法实际执行时间的资源因素是多方面的，如：网卡性能、 IDS 可支配的缓存量、总线调度方式、IDS 软件得到的 CPU 能力等。

定义 6　内存消耗指 IDS 占用的内存容量

$$f(\text{入侵特征规模，并发入侵数，数据到达强度，攻击密度})，$$

该测度反映 IDS 的空间性能。

入侵特征规模和并发入侵数对内存消耗的影响是显然的，包括数据到达强度和攻击密度参数是考虑到事件分析和报警生成时采用动态缓冲技术的可能性。

3.3 其它测度

通过功能和性能评估，人们可以了解一个 IDS 的有效性、效率和成本特征，因此这两者是目前 IDS 评估的主要内容。全面地描述 IDS 的系统能力，还应该包括下面测度：

定义 7　抗攻击能力指 IDS 抵抗针对自身的攻击的能力。

IDS 本身也会存在安全漏洞。若对 IDS 攻击成功，则直接导致其报警失灵，入侵者在其后的行为将无法被记录。因此 IDS 必须保证自己的安全性，这就需要考虑 IDS 本身的抗攻击能力。信息技术安全评估通用准则[17]（简称 CC）是被广泛应用于安全产品和系统的评估准则，如何将其应用于 IDS 是专门的研究内容。

定义 8　可用性指系统安装、配置、管理和维护的方便程度 $\sum_i k_i d_i$，其中 d_i 为每一项的得分，k_i 为权值。

有些文献提到日志、报警、报告以及响应能力[18]。日志、报警和报告是 IDS 对外输出的多种方式，方式越多，管理员使用 IDS 越方便，因此归属于可用性测度。而响应能力不在图 1 的抽象检测模型范畴。

4　评估的总体思路

评估的关键是定义 benchmark（基准的测试数据集）。用于系统能力评估的 benchmark 包括 3 类实例：入侵特征、入侵和正常行为实例，前者模拟安全专家的入侵知识，而后两者分别模拟实际环境中的入侵和正常行为。由于 IDS 表现出的漏报和误报是功能、性能以及其它可能因素综合的结果，为了计算某项测度值，benchmark 定义和评估的实施方案必须突出 IDS 在评估测度方面的差异，同时弱化其它因素的影响。因此理想的 benchmark 和评估的实施方案因测度不同而不同。

为了引用的方便，将 benchmark 中的入侵特征实例集记为 $C_{signature}$。由于入侵特征的描述能力不在系统能力的范围，因此不妨假定每个入侵特征实例既精确又完备，即所有与入侵特征实例匹配的事件序列构成入侵，所有与入侵特征不匹配的事件序列属于正常行为。那么，对 $\forall s1, s1 \in C_{signature}$，可逻辑地导出两个集合 $Intrusion(s1)$ 和 $NonIntrusion(s1)$，分别表示按照 $s1$ 分类的入侵空间和正常行为空间。这样，为每个测度定义 benchmark 的问题就转变为构造 $C_{signature}$，且对 $\forall s1, s1 \in C_{signature}$，计算 $C_{\text{int}\,rusion}(s1)$ 和 $C_{non\,\text{int}\,rusion}(s1)$，其中 $C_{\text{int}\,rusion}(s1) \subseteq Intrusion(s1)$ 表示入侵实例，$C_{non\,\text{int}\,rusion}(s1) \subseteq NonIntrusion(s1)$ 表示正常行为实例。

4.1　功能性测度

（1）benchmark 定义

为了充分地展示被测 IDS 在入侵特征分辨率方面的差异，评估入侵特征分辨率的 $C_{signature}$ 应覆盖所有维并且取最大维分辨率（如果被测 IDS 事先给出产品功能的描述，入侵特征实例只需要覆盖描述中给出的约束类型）。反之，如果是评估 IDS 的用户行为分辨率，入侵特征实例最好局限于所有 IDS 都能适应的范围，例如单事件特征。

在本文方法中，由于被测 IDS 统一配置评估方提供的入侵特征实例，同时入侵实例是入侵特征的实例化，因此被测 IDS 了解入侵实例的构成，完全可以通过配置虚假特征达到检测入侵实例的目的。例如 IDS 配置集合 $\{A, B, C, A \wedge B,\ B \wedge C, A \wedge C\}$ 的任意模式同样达到检测入侵 $A \wedge B \wedge C$ 的目标，其中大写字母是单事件模式，$\wedge$ 代表逻辑与。因此为了防止弄虚作假，评估不仅要考察 IDS 对入侵的识别能力，同时要考察它对正常行为的通过能力，当且仅当检测到入侵实例同时不生成对正常行为实例的误报才说明 IDS 适应被测入侵特征实例。即

$\forall s1, s1 \in C_{signature}$，$C_{\text{int}\,rusion}(s1)$ 是 $Intrusion(s1)$ 的代表子集，且 $C_{non\,\text{int}\,rusion}(s1)$ 是

$NonIntrusion(s1)$ 的代表子集。

Benchmark 定义的最后一项任务是将逻辑定义的入侵和正常行为实例映射为物理的审计数据流，映射方法同样和评估测度有关。对于入侵特征分辨率评估，我们理想化环境输入以弱化 IDS 的用户行为分辨能力的干扰，即：令一个审计数据包含一个逻辑事件的全部信息，这样无论 IDS 的实际采集能力如何，它总能正确分辨用户行为。反之，当评估用户行为分辨率时，benchmark 设计应充分考虑审计数据和逻辑事件的语义不一致性，例如将事件内容分解到多个审计数据中。

（2）评估的实施

评估时，被测产品统一配置评估方提供的入侵特征实例，然后在入侵和正常行为实例的作用下运行。当连续播放多个入侵或正常行为实例时，实例间彼此干扰，可能影响结果分析。最简单的方法是逐个实例进行测试，但这种方法效率低下。如何提高效率是今后研究的内容。

同时，评估系统应尽量慢地播放测试数据，以避免数据到达造成的压力。

（3）结果分析

漏报和误报对功能测度的评价有不同的权重。在本文方法中，IDS 了解入侵实例的构成，因此检测到一个入侵不代表该入侵的特征被正确检测；同时漏报一个入侵通常因为管理员少配置了相应特征，这种人为疏忽也不应影响对系统能力的评价。相反，误报明确说明 IDS 不适应测试方提供的入侵特征，或者配置了宽松的入侵检测规则或模式匹配算法不正确，因此误报对结论起决定性作用。尽管如此，漏报也不能被忽视。举个例子，评估方提供了 5 个入侵特征，要求被测 IDS 将它们配置到规则库中进行测试。假定某 IDS 弄虚作假，将它的规则库设为空，那么无论什么样的事件流通过，该 IDS 的误报数都为 0。因此，仅当漏报数为 0，误报数/误报率才描述了 IDS 的分辨率。

4.2 性能测度的评估

（1）benchmark 定义

有理由相信：对同一 IDS，当其执行不同复杂程度的功能时，负荷能力不同，因此性能测试必须首先明确对应的功能。功能决定了入侵特征实例的类型，即其必须覆盖的维以及维分辨率，同时压力条件（入侵特征规模）决定了入侵特征实例的数量，并最终决定了 $C_{signature}$。

同样，对 $\forall s1, s1 \in C_{signature}$，可确定两个集合 $Intrusion(s1)$ 和 $NonIntrusion(s1)$，但性能评估关于其中的哪些元素构成 $C_{\text{int }rusion}(s1)$ 和 $C_{non\,\text{int }rusion}(s1)$ 不作要求，只要这种构成对被测 IDS 是统一的，并且有足够的量造成压力。

（2）评估的实施

由于入侵特征实例由被测 IDS 配置到规则库中，被测系统可能弄虚作假，例如配置比入侵特征实例宽松的入侵检测规则或较少的规则数，从而换取高性能。为了阻止这种情况发生，在性能评估之前，应按照功能评估的方法验证规则库的配置。

其次，由于 IDS 的负荷能力是 IDS 的检测能力在拐点处的数据到达强度，那么即使测试实例之间存在干扰，这种干扰在计算相对值时已得到消除，不会引起拐点漂移。因此性能评估不要求对入侵/正常行为实例逐个测试，这使得性能评估具有可操作性。性能评估需要多次进行，每次设置不同的压力值，并递增数据到达强度，直到出现检测能力明显下降的拐点，该点处的数据到达强度就是 IDS 在压力条件下的负荷能力。同时考虑 IDS 的资源配置，就可得到处理效率。

（3）结果分析

原则上，IDS 的负荷能力通过检测能力（漏报和误报）随着数据到达强度的变化求得。但是当 IDS 面临压力，来不及处理而被丢弃的数据包是导致漏报和误报的真正原因，所以丢包可以归并漏报和误报两个维度的特征，并且测量和分析更加简单。在本文实验中，我们就是通过观察丢包而不是检测能力（漏报和误报）的变化计算负荷能力。

5 实验结果

5.1 合理性实验

为验证以系统能力为评估对象的合理性，本文对作者所在实验室开发的入侵检测系统 monster 和开放源码系统 snort[18] 进行比较。Monster 和 Snort 一样，都是基于单报文的 IDS，并且使用相同的规则描述语言。但 Monster 在报文采集模块和检测算法方面做了改进，目标是应用于高速网络。基于对这两个系统的实现的了解，合理的评价应是：monster 和 snort 功能相同，但前者具有更好的处理效率。

入侵特征的质量包括完备性和准确性两方面，因为只要展示了一方面对评估结论的影响就足以说明统一入侵特征的必要性，所以我们只进行完备性实验。实验使用 250 条随机抽取的 snort 规则，并根据 snort 规则构造 2000 条入侵报文作为测试数据。实验对 snort 和 monster 在不同流量速率下的检测能力进行评估，由于 Snort 只能工作在 100M 以下的网络流量环境中，因此两者只做 100M 流量以下的比较。所有条件下误报率为 0。

图 2 是当 snort 包含 250 条规则而 monster 只包含其中的一半数量时得出的检测率，数据表明：在小于 95M 的速率下 snort 的检测率高于 monster。综合检测率和误报率，传统方法将得出结论：小于 95M 的流量环境下，snort 的检测能力优越于 monster。图 3 是 snort 和 monster 包括同样的 250 条规则时得出的检测率，综合检测率和误报率，传统方法又将得出结论：在 100M 以内的流量环境下 monster 的检测能力优越于 snort。可见，传统方法对相同的两个系统得出不同结论。

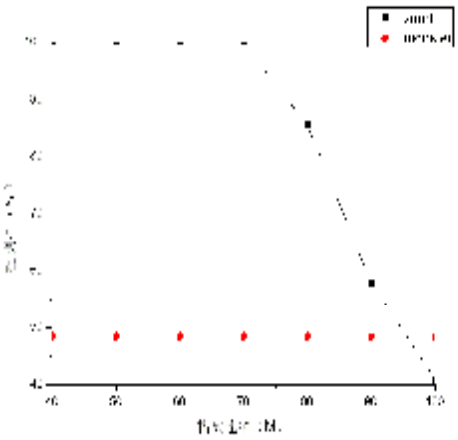

图 2 规则库不同时的检测结果

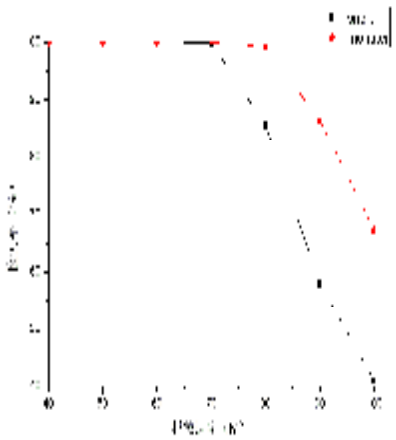

图 3 规则库相同时的检测结果

5.2 Snort 和 bro 的系统能力评估

Bro[19] 是另一个著名的开源 IDS。当只装载规则组件时，bro 与 snort 具有相似的功能特征，即允许用户将一些数据包特征定义成规则，然后检测软件（在 bro 中是事件引擎）根据规则匹配网络数据包生成相应的告警；然而与 Snort 简单粗糙的检测方式相比，Bro 支持正则表达式的数据匹配、规则关联、规则与策略脚本的交互，这些特性使得 Bro 具有准确检测的潜力。按照本文的术语，bro 的入侵特征分辨率大于 snort 的入侵特征分辨率（因为是开源软件，入侵特征分辨率可以通过对检测语言的分析得出）。

作为示例，实验对 snort 和 bro 基于单报文的检测效率进行测试。实验对象是关闭了所有预处理器和策略引擎的 bro 0.90。实验系统运行 Linux Redhat 9.0，硬件配置为 Pentium 4 Xeon 2.4G，双 CPU，2G 内存（bro 对硬件配置的要求是 2G 的 CPU 和 1G 的内存）。由于 bro 自己声明流量上限为 20kpps，因此我们只进行 20kpps 以下的测试。

实验分 2 组进行，第一组固定攻击比率为 1：20，规则数从 100 递增到 1000；第二组固定规则数为 1000，攻击比率从 1：1 递降到 1：10。由于是基于单报文的检测，不存在并发入侵的压力条件。

图 4 是 snort 和 bro 的负荷能力随规则数的变化趋势。数据表明，snort 和 bro 的负荷能力均随着规则数的增加而降低，但 bro 的降低趋势相对缓慢，这一方面表明规则数影响负荷能力，因此验证了规则数作为压力条件的合理性；另一方面说明规则数量对 bro 的性能影响小于对 snort 的影响。

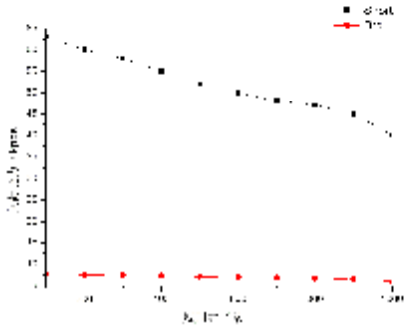

图 4 负荷能力随规则数的变化

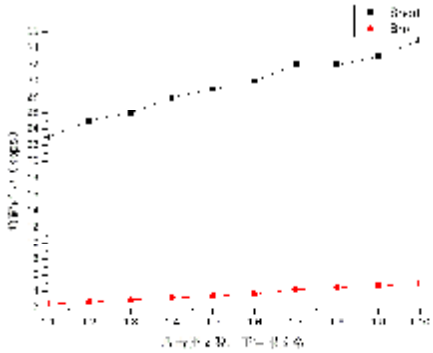

图 5 负荷能力随攻击比率的变化

图 5 反映 IDS 负荷能力随着入侵比率的变化趋势。随着入侵报文在测试数据中所占比重的降低，IDS 的负荷能力不断增加，表明入侵比率是合理的压力条件。同时，bro 的变化率略小于 snort，说明入侵比率对 bro 的影响小于对 snort 的影响。

综合图 4 和图 5，bro 在测试中始终表现出较小的负荷能力，原因是它的负荷能力基数小，而这个基数主要由事件采集模块的性能决定的。总之，尽管 Bro 在理论上做到了对入侵的准确检测，但它表现出的小负荷能力使得 bro 还难以作为一个实用的 NIDS 产品而得到推广。对 bro 的性能优化可重点关注事件采集模块。

6 结论及未来的工作

传统方法以 IDS 表现出的检测能力为评估对象，对基于特征的 IDS，评估结论是 IDS 实现和预置人工知识的综合质量。由于人工知识是独立于 IDS 的环境因素，将其纳入对 IDS 的评价是不合理的。在此基础上，本文给出基于特征的 IDS 的新评估体系，它以 IDS 的系统能力作为评估对象。本文重点研究了系统能力的测度选取，并分类讨论了测度计算的总体思路。实验表明：本文方法能够更真实地反映出 IDS 的质量，具有很好的实用性。

在评估的总体思路中，本文重点讨论各类测度的 benchmark 定义和评估实施方案的理想条件，但对如何满足条件没有讨论，这正是今后工作的重点。

参考文献

[1] Nicholas J.Puketza , et al. A Methodology for Testing Intrusion Detection System. IEEE Trans on Software Engineering,1996,22(10):719-729.

[2] Jesse C. Boothe-Rabek. WinNTGen: Creation of a Windows NT 5.0+ Network Traffic Generator [MS. Thesis]. MIT Department of Electrical Engineering and Computer,1998.

[3] Kristopher Kendall. A Database of Computer Attacks for the Evaluation of Intrusion Detection Systems [Bachelor Thesis].MIT Department of Electrical Engineering and Computer Science,1999.
[4] Kumar J. Das. Attack Development for Intrusion Detection Evaluation [Bachelor Thesis]. MIT Department of Electrical Engineering and Computer Science, 2000.
[5] 董晓梅,肖珂,于戈. 入侵检测系统评估技术研究. 小型微型计算机系统,2005,26(4):568-571.
[6] Alvaro A.Cardenas, John S.Baras, Karl Seamon. A Framework for the Evaluation of Intrusion Detection Systems. In: Proceedings of the 2006 IEEE Symposium on Security and Privacy.
[7] http://packetstorm.widexs.nl/UNIX/IDS/nidsbench/.
[8] http://osec.neohapsis.com/about.html.
[9] http://www.nss.co.uk/default.htm.
[10] Hilmi Gunes kayacik. The Chanllenges in Traffic and Application Modeling for Intrusion Detection System Benchmarking. In : Proceedings of the RAID International Symposium, 2004.
[11] 钱俊,许超,史美林. 入侵检测系统评测研究进展（上）.计算机安全,2005,8:17-20.
[12] 钱俊,许超,史美林. 入侵检测系统评测研究进展（下）.计算机安全,2005,8:16-17.
[13] John Mchugh. Testing Intrusion Detection Systems: A Critique of the 1998 and 1999 DARPA Intrusion Detection System Evaluations as Performed by Lincoln Laboratory. ACM Transactions on Information and System Security, 2000, 3(4): 262-294.
[14] R. Davis, H. Shrobe, and P. Szolovits. What is a knowledge representation? .AI Magazine, 14(1):17-33, 1993.
[15] H. Debar, et al. An Experimentation Workbench for Intrusion Detection Systems. Research Report,RZ2998, Switzerland Petri Mahonen :IBM Zurich Research Laboratory,1998.
[16] S. Kumar. Classification and Detection of Computer Intrusions[PhD Thesis]. Dept. of Computer Science, Purdue University,USA, 1995.
[17] National Standard of China. Evaluation criteria for information technology security (GB/T 18336-2001). 2001,12.
[18] Sourcefire.Snort 2.0. http://www.sourcefire.com/technology/whitepapers.htm.
[19] Vern Paxson, Jim Rothfuss, brian Tierney. Bro User Manual. http://bro-user-manual.pdf/,2004-12-1.