入侵检测系统评估方法综述

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

A Survey of the Approaches of Intrusion Detection System Evaluation

Wang Yang , Gong Jian
（Dept. of Computer Science and Engineering, Southeast University, 210096 Nanjing）
（Computer network technology key laboratory, Jiangsu）

【摘要】本文对入侵检测系统评估方法的发展进行了综述，讨论了评估入侵检测系统的目的和难点，对 90 年代中期至今出现的各种评估方法进行了介绍。文章着重分析了 Nicholas J.Puketza 等人的工作和 DARPA 的评估工作，和这些评估方法的不足，指出进行大规模分布式网络入侵检测系统的测试是目前研究的热点。
【关键词】入侵检测系统（IDS），评估，测试平台
【中图分类号】TP393

【Abstract】This paper is a summary of the development about intrusion detection system evaluation. With indicating the purpose and difficulty of intrusion detection system evaluation, this paper introduces several approaches of evaluation from 1990's until now. Nicholas J.Puketza's work and DARPA's evaluation with their shortage are introduced in details. This paper also indicates that evaluation of intrusion detection system on large scale network is the trend today.
【Keywords】Intrusion Detection System(IDS), Evaluation, test bed

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，因此对各种入侵检测系统的功能和性能评估也形成了需求。最早的 IDS 系统通过审查系统日志来发现特定主机上的异常用户行为，这类 IDS 系统是基于主机的。随着入侵检测技术的发展，1990 年出现了通过网络监听进行入侵检测的系统，1991 年出现分布式入侵检测体系。20 世纪 90 年代中后期入侵检测的体系结构的研究又有了许多发展，例如 Mark Crosbie 和 Gene Spafford 的自治代理概念被广泛接受【12】。 对于 IDS 系统的测试也相应处于不断发展之中，1996 年 Nicholas J.Puketza 等人提出通过模拟被监测系统上用户的行为来测试 IDS 系统的方法【1】。1998、1999 年 MIT 的 Licoln Laboratory 通过模拟美军空军基地的日常网络流量进行了 IDS 系统的两次测试。2000 年时，MIT 在测试数据中加入了对于 DDoS（Distributed Deny of Service）攻击的模拟。2001 年 Terrence Champion 等人专门对于 IDS 系统检测 DDoS 攻击的能力进行了测量，同时还考虑了 IDS 系统响应功能的评测【10】。与此同时，一些商业厂家也给出了自己的入侵检测系统的评估方法。

本文首先给出了评估工作的目的和难点，然后着重介绍了 Nicholas J.Puketza 等人和 MIT 的评估方法并指出了这些评估方法的不足。接着介绍了一些商业厂家的评估方法和目前国内研究状况。

基金项目：国家自然科学基金课题资助(90104031)

2. 评估的目的和难点

现有的对入侵检测系统评估研究的目的在于：(1) 给出对特定 IDS 系统的功能、性能的评估。使用入侵检测系统的组织必须知道自己使用的入侵检测系统的工作效率如何，并且由此决定对此系统的依赖程度的大小，以及是否需要使用其他的安全措施。(2) 通过对测试结果的分析，改进入侵检测的算法。(3) 总结现有算法的不足，发现新的入侵检测算法。(4) 探讨一种统一的测试方法和评估标准。

评估入侵检测系统存在的难点为：①鉴别一个特定 IDS 系统将遇到的攻击集是十分困难甚至是不可行的。原因有三：首先目前已知的攻击技术的总数目十分巨大；其次一个特定的 IDS 系统不会遇到过去发现的所有攻击；最后攻击者会找寻系统新的漏洞，并且以此构造新的攻击方法【1】。②IDS 的工作效率将受到其所在的网络中众多条件的影响【1】。③没有统一的测试方法和评估标准。虽然 UC Davis 分校提出了 CIDF(The Common Intrusion Detection Framework Architecture) 【11】，IETF 的 IDWG 也提出了相应的安全检测框架模型，但是到目前为止没有入侵检测系统的公认框架标准，所以要针对实现各异的入侵检测系统制定统一的测试方法和评估标准十分困难。至今没有研究机构给出一种适用的评估框架，因此对于各种入侵检测系统没有办法做出有效的横向比较。④构造测试所需的网络环境、背景数据、攻击数据以及对测试结果进行分析是耗时和工作量巨大的。

3. Nicholas J.Puketza 等人的工作

Nicholas J.Puketza 等人的测试方法针对当时已有的 IDS 系统，提出检测范围、检测的系统开销、高负荷下的检测能力是三项通用评估指标，并对应三项指标进行了入侵鉴别测试、资源使用率测试、高负荷测试。每项测试包含两个步骤：测试用例选择和测试过程。测试用例的选择基于分类的方法，对已有攻击进行分类，然后从每一个类别中挑选一个攻击用于测试。测试在一个独立的局域网上进行，使用 unix 平台上的 expect 和 tcl 来运行一段脚本以模拟一个计算机的用户，同时运行多个脚本就可以模拟出多个用户使用多台计算机的效果。通过组合不同的脚本，可以同时模拟正常用户的行为和入侵者的行为。测试过程通过一些精心设计的测试场景来实现，每一个测试场景均由下面的基本测试过程组合而成：(1) 创建或者选择测试脚本集。测试脚本集中的脚本描述了用于本次测试的正常用户行为和攻击行为。(2) 建立必要的测试环境，如创建测试环境中计算机之间系统级的底层通讯。(3) 启动 IDS。(4) 运行测试脚本。(5) 分析 IDS 的输出。通过对照 IDS 的输出与模拟攻击行为的脚本可以给出被测 IDS 的测准率。通过测试过程中纪录的资源使用状况可以给出 IDS 的资源使用率。Nicholas J.Puketza 等人运用上述测试方法对 UC Davis 开发的 Network Security Monitor(NSM) 进行了测试，测试结果表明该测试方法能够提供和 IDS 性能相关的有用信息。但即使是针对当时的 IDS 系统而言，该测试方法还存在一些缺陷：(1) 不能够完全模拟使用图形界面用户的行为。(2) 测试方法是为使用滥用检测算法的 IDS 设计的，对于使用异常检测算法的 IDS 不适用。Nicholas J.Puketza 等人的方法的要点是设置了一组各不相同的 IDS 通用的性能目标，使得测试工作不再是针对某个特定系统。

4. MIT 的工作

MIT的林肯实验室（Lincoln Laboratory）在DARPA资助下所开展的IDS评估工作是目前国际上最为领先和完整的。他们在1998年为研究入侵检测系统而进行了第一次系统性的实际测试，为参加测试的6个系统了提供封闭的测试环境和测试的数据集，通过黑盒测试得到测试结果。对测试结果的分析方面，MIT通过使用接受者运行特征(ROC)曲线来揭示被测系统测准率和误报率的函数关系。MIT的测试数据模拟了一个美军空军基地的日常网络流量，通过使用由AFRL(Air Force Research Laboratory)修改过的Unix内核，一个具有十几台工作站

基金项目：国家自然科学基金课题资助(90104031)

的网络可以模拟出一个具有超过1000台工作站和100个用户的网络。这个网络具有两个逻辑部分：“内部”（例如，基地内部）和“外部”(例如：Internet)，通过路由器实现互联。林肯实验室通过这个独立的计算机网络制造了适应性数据和测试数据，其中适应性数据和测试数据具有相同的流量特征，但是适应性数据只含有已知的攻击数据，而测试数据中具有新的攻击数据。适应性数据用来对使用异常检测算法的系统进行训练，而测试数据用于正式测试。98年的攻击类型主要集中在UNIX、路由和新的攻击方面，测试的结果表明各个被测系统对于探测(probe)和本地超户权限窃取（u2r）攻击的检测情况比较好，而对于远程获得本地权限（r2l）和新型的DoS攻击的检测情况比较差【3】。98年的测试结果表明下一步的入侵检测研究不能只基于通过使用攻击特征来检测攻击的方法，而是应该研究能够发现新攻击的算法。

1999年有8个系统参加了测试，其中包括了使用异常检测算法的系统。99年MIT的测试沿用了98年的框架，但它涵盖了更多的攻击类型。99年测试加入的新特征为：①加入了NT工作站作为被攻击对象，背景数据中有了NT产生的数据，攻击数据集中增加了对于NT的攻击。②增加了内部攻击。③不再提供被攻击主机上全部文件系统的dump数据，只提供文件系统重要部分的信息（包括NT的审计日志）。④提供内部网络上窃听（sniffing）到的数据。⑤提供了被模拟的空军基地的安全政策。另外99年的测试对攻击数据按类型进行了划分，这样不同的IDS系统可以选择不同的测试数据集，在分析测试结果时也可以针对各类数据集给出ROC曲线。99年的测试简化了评分方法，用于解决98年在分析测试结果时遇到的困难。99年的测试结果表明：基于网络的入侵检测系统对于老的探测和老的DoS攻击的检测率较高；基于主机的入侵检测系统对于Solaris的user-to-root(U2R)攻击的检测率较高；将基于网络和基于主机的入侵检测系统结合使用可以提供最好的综合检测效果。但是被测系统对于从未出现过的攻击、隐秘攻击和Windows NT攻击的检测率很低，58种攻击中有10种没有被参加测试的任何一个系统检测出来，原因在于：①没有一个系统对协议和TCP服务进行了深入地分析；②已知攻击的特征不能为检测新攻击提供有用信息；③并不是每台计算机都能提供有效的日志。 测试结果也表明基于主机的入侵检测系统、采用异常检测算法的入侵检测系统、对于文件系统进行深入分析的入侵检测系统依然有很大的发展前途【6】。

2000 年 MIT 测试的目的是为了帮助 IDS 的开发者改进系统，而不是评估系统性能。MIT 在攻击数据中加入了 DDoS(Distributed Deny of Service)攻击，希望通过提供新的攻击帮助开发者改进算法、建立更好的入侵检测模型。

MIT的离线测试只是DARPA的入侵检测系统评估计划的一部分，另一部分则是98年和99年AFRL对参加评估的一部分入侵检测系统进行的实时测试。测试使用的是AFRL的试验床，被测系统必须实时检测攻击。MIT指出在线测试的优点是：①实时测试中，被测系统可以根据实际情况做出响应。②实时测试可以检测入侵检测系统的资源使用率和安装配置的难易程度。缺点是：实时测试需要对每个被测系统重复一次，因此十分的耗时，也不易调整来为异常检测系统作训练【5】。

针对98，99年测试的缺点，综合实时测试的优点，MIT提出了一些改进方案【14】。第一、改进测试使用的测试平台，建立了LARIAT (the Lincoln Adaptable Information Assurance Real-time Testbed)。LARIAT包括了测试用的测试平台和一组图形界面的软件工具。利用LARIA，IDS的研究和开发者可以在自己的实验室中进行自定义的实时测试。在实际测试中，LARIAT既能够产生和控制背景流量，又能够按预先配置的脚本产生攻击数据。第二、产生并发布了两个特定场景的数据集。和98和99年测试中使用的数据集不同，特定场景的数据集只包括真实的背景数据和一种特定的DDoS攻击的各个步骤：扫描（scanning），探测（probing），闯入（break-in），安装（installation）和发起攻击（launching of DDoS）。第三、给出了一种对攻击特征和IDS特征进行统一描述的模式化方法，利用这种模式化的描述

基金项目：国家自然科学基金课题资助(90104031)

可以直观地给出IDS的特征，从而确定测试需要的数据集。

2001 年 AFRL 的 Terrence Champion 专门对网络入侵检测系统进行了一次测试，测试的目标是对网络入侵检测系统检测分布式攻击的灵敏性进行量化评估。参加这次测试的有两个系统：一个是采用滥用检测算法的商用系统，该系统依靠给定的攻击特征进行检测。一个是 DARPA 提供的异常检测系统，该系统采用统计方法进行检测。Terrence Champion 使用参加攻击的主机数目和每个攻击动作的时间间隔作为参数来描述 DDoS 攻击的强度，给出了在各种参数组合下被测系统的检测结果。测试的结论是：漏过检测的攻击能够对被攻击主机的性能产生明显的影响。DARPA 的系统比商业系统有更强的检测能力，同时能够为响应设备提供足够的信息【10】。

5. 其他人的工作

一些商业厂家也给出了自己的测试方法。IBM 给出了一种基于规则的测试方法【8】，这种方法使用一组有代表性的动作集来测试 IDS 的误报率和漏报率；并希望通过测试得出一种 IDS 的组合使用方法，使得综合漏报率和误报率最低。评估使用的方法是对网络活动中的恶意动作进行形式化的描述得到一组有害动作集，对于 IDS 系统的检测能力进行形式化的描述得到检测能力集，对于每个有害动作触发的报警信息进行形式化的描述得到报警信息集。将有害动作集和 IDS 的检测能力集做与运算可以得出该系统应该能够检测出的有害动作集，并由此得到相应的报警信息集。将实际测试的结果与按规则进行运算得到的结果进行比较可以确定被评估系统的性能。IIS（Internet Security System）公司没有给出具体的评估方法，但是给出了 IDS 应该达到的一组评估标准：IDS 应该能够理解网络会话的语义，而不会被同一信息的不同表示欺骗。IDS 应该配置灵活且简易，具有多种报警方式和良好的性能。在不断增加攻击描述的情况下，IDS 仍旧能够正常工作【9】。

目前国内对于入侵检测系统的评估工作还处在起步阶段，主要工作是：一、对已有的评估方法进行重复、模拟和改进。二、根据实际使用经验提出 IDS 系统应该满足的指标。参考文献【15】就给出了一组详细的测试指标。通过指标来衡量的方面包括：基本特性、功能、报告和审计、检测和响应、安全管理、安装和维护、配置和售后服务。

6. 综合评价

Nicholas J.Puketza 等人的工作借鉴了软件测试学中的方法，给出了一组通用的评估指标和测试场景。测试中使用的关键技术是建立了一个通用的软件平台，在该平台上能够通过不同的脚本模拟出不同的用户行为。该方法的缺点是：1、软件平台只能覆盖小范围的局域网，对于大量用户的模拟需要复杂的软件结构和巨大的计算能力。2、随着各种网络应用的不断出现，用户的行为模式越来越复杂，对于用户行为的描述需要越来越复杂的脚本支持。

MIT 的评估工作是系统和通用的。全部工作包括数据集的构造，测试过程，测试结果的分析，提供改进算法的意见。评估的对象包括各种类型的入侵检测系统。MIT 为 IDS 的评估提供了一组通用数据集和评估方法。从 MIT 测试方法的发展可以看出：由采用静态数据集的离线测试逐步发展为采用可根据实际情况进行配置的实时测试；由采用统一的数据集逐步发展为采用针对特定系统和特定场景的数据集。MIT 的测试存在以下问题：1、测试得到的结论只能被放在测试用的试验网，攻击数据，背景流量和评分方法等相关上下文中理解。测试使用的网络拓扑简单，背景流量很小，攻击数量有限。2、构造测试使用的试验网，背景数据和攻击数据集是极其耗时和工作量巨大的工作，对于测试结果的分析也需要大量的人力。所以 98 和 99 年的每次测试都耗时一年。因此从使用大而全的测试数据集向使用针对特定系统和特定场景的测试数据集转化有利于降低测试工作的复杂度。

基金项目：国家自然科学基金课题资助(90104031)

7. 结论

网络的发展和新攻击的不断出现推动了入侵检测系统的发展和完善，入侵检测系统的测试技术也在不断进步。从 90 年代中期，测试基于主机的入侵检测系统，发展为测试基于网络的入侵检测系统，测试混合型入侵检测系统；目前的发展方向是测试具有协同和响应功能的大规模分布式网络入侵检测系统。测试所使用的方法也从提供主机的审计文件发展到提供模拟网络流量的静态数据集，由离线的黑盒测试发展为在线的实时测试。测试平台应该能够提供各种特定场景和攻击数据，具有模拟大规模网络流量和分布式攻击的能力。目前的测试工作还侧重于滥用检测系统的测试。对于采用异常检测算法，免疫学算法的新式系统的测试方法还有待完善。

参考文献

[1] Nicholas J.Puketza , et al. A Methodology for Testing Intrusion Detection System[J] IEEE Trans on Software Engineering,1996 22(10):720-728.

[2] Richard Lippmann, et al. 1998 DARPA Intrusion Detection Evaluation Plans: Part 1[R]. MIT Lincoln Laboratory, 1998

[3] Richard Lippmann, Robert K. Cunningham, et al. Results of the DARPA 1998 Offline Intrusion Detection Evaluation[R]. MIT Lincoln Laboratory, 1999

[4] Richard Lippmann, et al. Proposed 1999 DARPA Off-line Intrusion Detection Evaluation Plans[R]. MIT Lincoln Laboratory, 1999

[5] J.W.Hains, Richard Lippmann, et al. 1999 DARPA Intrusion Detection Evaluation: Design and Procedures[R]. MIT Lincoln Laboratory, 2001

[6] Richard Lippmann, et al. The 1999 DARPA Off-Line Intrusion Detection Evaluation[R]. Lincoln Laboratory MIT, 244 Wood Street, Lexington, MA 02173-9108 , 2001

[7] Terrence G. Champion, Robert S. Durst Air Force Intrusion Detection System Evaluation Environment[R] Air Force Research Laboratory,1999

[8] Dominique Alessandri. Using Rule-Based Activity Descriptions to Evaluate Intrusion-Detection System[R]. Switzerland, IBM Research Laboratory Zurich, October 2000

[9] The Evolution of Intrusion Detection Technology[Z].ISS white paper

[10] Terrence Champion. A benchmark evaluation of network intrusion detection systems[J]. Aerospace Conference, 2001, IEEE Proceedings.

[11] The Common Intrusion Detection Framework Architecture[S].

[12] 龚俭,陆晟. 大规模互联网络的入侵检测[J]. 东南大学学报（自然科学版） 2002 年 5 月

[13] 张宇,贲可荣. 美国国防部入侵检测评估综述[J].通信技术,2002

[14] Joshua W. Haines, Lee M. Rossey, et al. Extending the DARPA Off-Line Intrusion Detection Evaluations[R]. Lincoln Laboratory, Massachusetts Institute of Techonology, 2002

[15] 罗嵘. 入侵检测产品的评价指标[J].通信技术,2001

基金项目：国家自然科学基金课题资助(90104031)