蒋盛益

基于聚类的入侵检测算法研究

科学出版社
www.sciencep.com

基于聚类的入侵检测算法研究

蒋盛益

科学出版社

北京

内容简介

本书以聚类分析为基本工具，围绕入侵检测这一目标展开讨论。包括：入侵检测的相关概念和研究现状；聚类分析的基础知识；针对入侵检测问题中数据具有大规模和混合属性的特点，重点研究数据之间的差异性度量方法，高效自适应聚类算法；在介绍现有经典的异常挖掘算法的基础上，提出异常因子的度量方法，进而研究静态异常挖掘算法；改进最近邻分类方法，在静态异常检测的基础上，研究一类动态数据的异常检测，将入侵检测问题视为动态数据的异常检测问题；总结并对后续研究工作进行展望。

本书通过实例说明原理，对从事数据挖掘、入侵检测的科技人员具有重要的参考价值。还可作为计算机、信息技术等专业的研究生学习、研究的参考资料。

图书在版编目(CIP)数据

基于聚类的入侵检测算法研究/蒋盛益. —北京：科学出版社，2008

ISBN 978-7-03-022553-5

Ⅰ. 基… Ⅱ. 蒋… Ⅲ. 计算机网络-安全技术-研究 Ⅳ. TP393.08

中国版本图书馆 CIP 数据核字(2008)第 108149 号

责任编辑：王雨舸 / 责任校对：曾 莉

责任印制：董艳辉 / 封面设计：苏 波

科学出版社 出版

北京东黄城根北街 16 号

邮政编码：100717

http://www.sciencep.com

武汉市科利德印务有限公司印刷

科学出版社发行 各地新华书店经销

*

2008 年 8 月第 一 版 开本：A5(890×1240)

2008 年 8 月第一次印刷 印张：7 3/4

印数：1—1 500 字数：240 000

定价：28.00 元

(如有印装质量问题，我社负责调换)

前　言

随着网络和其他信息技术的广泛应用，全球信息化的步伐越来越快，网络信息系统已成为一个单位、一个行业，乃至一个国家持续发展的基础设施，因此，网络安全也就成为国家与国防安全的重要组成部分，网络信息安全也就成为影响国家长远利益和持续发展的重大问题。入侵检测技术是一种重要的动态安全防护技术，已成为信息技术的一个重要研究方向。由于数据挖掘技术，可以从海量审计数据中发现正常和异常的行为模式，不仅可以减少人工分析和编码带来的繁重工作，而且可以提高入侵检测系统的适应性，因此近年来数据挖掘的各种技术被大量应用于入侵检测。

聚类分析是数据挖掘的重要手段，特别适合于数据点之间的内部关系的探索，广泛应用于一些探索性领域。从数据处理的角度来看入侵检测，入侵行为对应的数据可以看成异常数据，因而入侵检测问题也就可以看成一种特殊的异常数据挖掘问题。通常，可以假设正常行为和入侵行为具有不同的特征，是可以区别的，入侵行为偏离正常行为，在某种特征空间中，正常行为与入侵行为对应的数据会彼此分离，而相同类别行为(正常或入侵)对应的数据彼此接近(即聚合在一起)。基于这样的考虑，导致了基于聚类分析的异常数据挖掘和入侵检测方法的研究。

本书以聚类分析为基本工具，围绕入侵检测这一目标展开讨论。全书共6章。第1章介绍入侵检测的相关概念和研究现状。第2章

讨论聚类分析的基础知识，介绍了典型聚类方法的思想，相似性度量及聚类算法的性能评价等内容。第3章针对入侵检测问题中数据具有大规模和混合属性的特点，重点研究了数据之间的差异性度量方法，高效自适应聚类算法。第4章介绍了现有经典的异常挖掘算法，重点研究了数据、数据子集偏离整体数据程度的度量方法，进而研究静态异常挖掘算法。将最近邻分类方法进行改进，在静态异常检测的基础上，研究动态数据的异常检测，将入侵检测问题看成动态数据的异常检测问题，这是第5章研究基于聚类的入侵检测方法的基础。第6章对书中提到的主要研究工作进行总结，并对后续研究工作进行了展望。

本书是作者相关研究工作的总结，主要工作是在华中科技大学国家高性能计算中心（武汉）读博士期间完成的。这些工作得到了国家自然科学基金项目“面向网络入侵检测的并行数据挖掘技术研究”、“面向数据流的异常挖掘算法研究”，广东省高等学校自然科学研究重点项目“数据流异常挖掘及在欺诈检测中的应用研究”的支持。本书的出版得到了广东外语外贸大学信息科学技术学院和科学出版社的大力支持，同时书中参考了许多学者的研究成果，在此一并表示衷心感谢。

限于作者学识水平，书中可能存在不足和疏漏，敬请同行和读者批评指正。

蒋盛益

2008年4月

目　录

第 1 章　入侵检测概述

1.1　研 究 背 景

中国互联网络信息中心(CNNIC)于 2008 年 1 月 17 日发布了第 21 次《中国互联网络发展状况统计报告》[1],报告显示,截止到 2007 年 12 月 31 日,我国的上网计算机总数达到了 7800 万台,同上一次调查结果相比,我国的上网计算机总数半年增加了 1090 万台,增长率为 16.2%,和上一年同期相比增长 31.3%,是 1997 年 10 月第一次调查结果 29.9 万台的 260.9 倍,图 1-1 显示了 2000 年 12 月以来历次调查上网计算机总数的情况。可见我国上网计算机总数呈现出比较快的增长态势。截至 2007 年 12 月,我国网民数已增至 2.1 亿人,略低于美国的 2.15 亿,位于世界第二位。中国网民数增长迅速,比 2007 年 6 月增加 4800 万人,2007 年一年则增加了 7300 万人,年增长率达到 53.3%,即平均每天增加 20 万人,是 2000 年 12 月的 9.3 倍。图 1-2 显示了 2000 年 12 月以来我国历

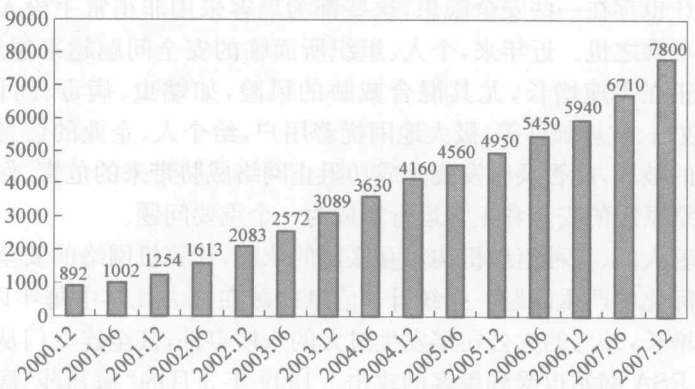

图 1-1　历次调查上网计算机总数(万台)

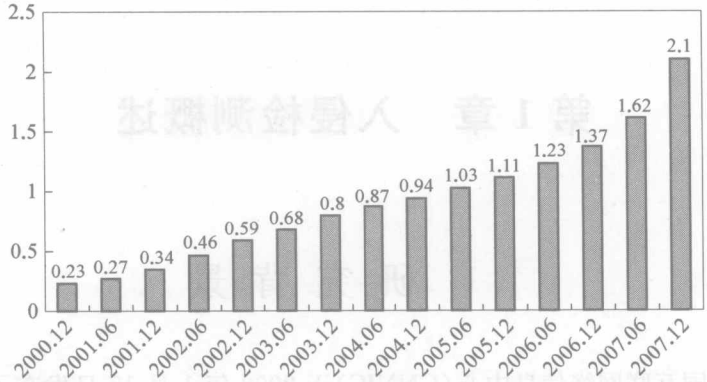

图 1-2 历次调查网民人数(亿人)

次调查网民人数的情况。2008 年 2 月 23 日,新华社发布消息称,中国网民数已达 2.21 亿人,超过美国居全球首位。

随着计算机和网络技术的发展,计算机网络已成为社会生活不可或缺的一部分,电子商务、电子政务、虚拟社区等建立在 Internet 网络上的电子在线服务呈快速增长的趋势,人类社会对数字信息的依赖达到前所未有的程度。信息的获取和交流越来越方便,由于 Internet 所具有的开放性和共享性在给人们带来便利的同时也带来了更多的安全隐患,对信息安全问题提出了严峻的挑战。由于系统安全脆弱性的客观存在,操作系统、应用软件、硬件设备不可避免地会存在一些安全漏洞,网络协议本身的设计也存在一些安全隐患,这些都为黑客采用非正常手段入侵系统提供了可乘之机。近年来,个人、组织所面临的安全问题越来越复杂,安全威胁正在飞速增长,尤其混合威胁的风险,如蠕虫、病毒、间谍软件、DDoS 攻击、垃圾邮件等,极大地困扰着用户,给个人、企业的信息网络造成严重的破坏,能否及时发现并成功阻止网络威胁带来的危害、保证计算机和网络系统的安全和正常运行便成为一个重要问题。

随着人们、各种组织机构对互联网的依赖,计算机网络的安全已成为一个国际化的严重问题。据统计:信息窃贼在过去几年中每年以 250% 的速度增长,99%的大公司都发生过大的入侵事件,甚至连专门从事网络安全的 RSA 网站也受到黑客的攻击。1999 年 3 月的"梅利莎"病毒事件导致电子邮件系统瘫痪,并造成了 8000 万美元的经济损失。据美国 FBI

的调查，美国每年因为网络安全造成的经济损失超过170亿美元。1999年12月不法分子从网上唱片零售商“CD宇宙公司”窃取了30万张信用卡的号码。2000年2月8日至10日，Yahoo、EBay、Amazon等8家美国著名大型网站相继遭到黑客袭击，使这些网站在相当长的时间内服务中断，网上交易完全陷入停顿，据估计，这次大范围网站瘫痪造成的经济损失和维修费用超过12亿美元。2003年1月25日，互联网遭到全球性的病毒——蠕虫王 Worm Netkiller 2003攻击，亚洲、北美和欧洲的网络陷入了瘫痪或者半瘫痪的状态，全世界范围内损失额高达12亿美元。美国NIFWIC (National Internet Fraud Watch Information Center)的数据显示，2005年美国由网络欺诈造成的总损失达到1386.3万美元，较2004年增长139.6%，平均损失也从2004年的895美元增长到2005年的1917美元，增长了114.2%。

中国从1986年首例计算机犯罪被发现至今，涉及计算机网络的犯罪逐年大幅上升，1999年公安机关立案侦查的计算机违法犯罪案件仅为400余起，2000年剧增为2700余起，2001年达到4500余起，2002年增加到7000起，绝大部分计算机违法犯罪案件牵涉网络[2]。2004年全国信息网络安全状况调查结果显示，在被调查的7072家政府、金融证券、教育科研、电信、广电、能源交通、国防和商贸企业等部门和行业的重要信息网络、信息系统使用单位中，发生网络安全事件的比例为58%。据中国国家计算机网络应急技术处理协调中心抽样统计：中国有超过250万台电脑被黑客用“僵尸网络”控制，用来向全世界发送垃圾邮件；仅2004年上半年，我国遭到Mydoom蠕虫、利用RPC漏洞和LSASS漏洞的几类主要蠕虫攻击的主机数目接近200万台；有70万台电脑被植入间谍软件，黑客可以随时从中盗走各种资料；2.2万多台电脑被木马病毒控制，计算机上的全部信息都暴露在黑客面前。2006～2007年，无论是网络仿冒、网页恶意代码还是网站篡改，增长的速度都达到150%～200%，木马主机年增长率更是达到2125%。2007年监测到中国内地有90万个IP地址主机被植入木马，比2006年增加21倍。内地有11万台主机作为控制端控制了我国被植入木马的计算机，内地有360万个IP地址主机被植入僵尸程序。根据我国公安机关发布的数据，2005年我国共处理网络安全犯罪近3万起，银行等金融机构成为网络诈骗的“重灾区”，造成的直接经

济损失约10亿元人民币[3]。

随着网络技术的飞速发展和Internet的日益普及，信息安全的重要性和必要性也愈加凸显，网络安全问题已经引起各国、各部门、各行各业以及每个计算机用户的充分重视，据市场研究公司Input发表的研究称，联邦机构IT安全开支将稳步增长，到2008年将达到60亿美元。计世资讯(CCW Research)研究数据显示，中国网络安全产品市场增长快速，2005年中国安全产品市场规模达到36.5亿元，比2004年同比增长27.3%。预计在2006～2009年的复合增长率为22.1%，2009年中国安全产品市场规模将比2005年翻一番以上，达到81亿元人民币[4]。

我们已跨入信息时代，信息已成为社会发展的重要战略资源，社会的信息化已成为当今世界发展的潮流和核心，而信息安全在信息社会中扮演着极为重要的角色，它直接关系到国家安全、企业经营和人们的日常生活。信息安全已成为国家与国防安全的重要组成部分，同时也是国家网络经济发展的关键。在目前信息安全形势严峻的情况下，大力发展信息安全技术，使日益增加的计算机及网络犯罪受到应有的制裁，进一步保护国家的安全不受侵犯，保障国家的经济秩序不被破坏，保护网络用户的合法权益不受侵害，具有非常重要的现实意义。可以说信息安全问题已成为国家信息化发展道路上的瓶颈，不从根本上彻底解决信息安全问题，信息化也不可能深入发展。从信息安全角度看，信息化程度越高，国家安全、社会安全面临的风险越大。信息安全问题已经成为制约我国经济社会发展的关键问题之一。因此对入侵攻击的检测与防范、保障计算机系统、网络系统及整个信息基础设施的安全已经成为刻不容缓的重要课题。

入侵检测技术是继防火墙、数据加密等传统保护措施后新一代的网络安全保障技术[5]，是一种主动的安全防护技术，它从系统内部和网络中收集信息，从这些信息中分析、判断系统是否存在安全问题，并采取相应的措施。入侵检测的前提是用户和程序的行为可以被监控，且正常行为和攻击行为是可以区分的。理论上讲，任何入侵行为都能够被发现，因为网络流量和日志文件记录了所有的网络活动。

评价一个入侵检测系统的质量有许多方面，有效性、适应性、时效性是三个重要的因素。有效性是指入侵检测系统具有高的检测率和低的误报率；适应性是指系统不仅可以检测已知攻击，还能检测到这些已知攻击

的变种或新的攻击;时效性是指对待检测数据能在有效的时间内得到检测结果。由于网络流量数据十分庞大、更新快,攻击行为一般很难被直接、及时发现,而现行的大多数入侵检测系统主要是利用手工获得的专家知识和经验来检测入侵行为,不仅对新的攻击或特征未知的攻击无能为力,而且检测的准确性与时效性达不到实际应用的要求。因此,用数据挖掘方法提取入侵特征、建立检测模型,无疑是实现入侵检测自动化的一条重要途径。近年来,基于数据挖掘的入侵检测技术得到了长足的发展,但现有基于数据挖掘的入侵检测方法在这三个方面尚不能满足实际应用的要求。

对于入侵行为的识别,关键是对被监控系统或用户的行为进行分类。通常可以假设,正常行为和入侵行为具有不同的特征,是可以区别的,入侵行为偏离正常行为,在某种特征空间中,正常行为与入侵行为对应的数据彼此分离,而相同类别行为(正常或入侵)对应的数据彼此接近(即聚合在一起)。入侵检测过程可看成数据分析的过程,入侵行为对应的数据可以看成异常数据,因而入侵检测也就对应异常数据挖掘。基于这样的考虑,将聚类分析技术、异常数据挖掘技术应用于入侵检测,研究基于聚类的入侵检测算法,并围绕如何提高检测算法的有效性、适应性和时效性进行研究。

1.2 计算机安全与入侵检测

计算机安全的广义定义是:主体的行为完全符合系统的期望,系统的期望表达成安全规则,也就是说主体的行为必须符合安全规则对它的要求。计算机信息系统安全包括以下几个方面的内容:

- **机密性**(confidentiality) 防止系统内信息的非法泄漏,即使信息不泄露给非授权的个人、实体或进程,不为其所用。
- **完整性**(integrity) 防止系统内软件与数据被非法篡改或破坏。
- **可确认性**(accountability) 当计算机遭受攻击后,安全系统有充足的信息追踪和识别入侵者。
- **可用性**(availability) 要求信息和系统资源可以持续有效,而且授

权用户可以随时随地存取资源。

上述计算机安全的各项特性取决于计算机系统安全策略的需求，这些安全策略用来定义或描述系统的不同用户和软件模块的行为，并明确指出哪些行为是合法的，哪些是被禁止的。一个安全的计算机信息系统应达到以上目标，为此，人们投入很多精力来开发信息安全技术，以保障信息安全。

1.2.1 被动安全防护技术

传统的信息安全技术，大都集中在系统自身的加固和防护上，属于被动的安全防护技术，这些安全技术主要有：密码技术、防火墙、鉴别与认证、访问控制、虚拟专用网等。然而，在被动安全防护技术下完全安全的信息系统根本不可能实现，被动安全防护技术主要存在以下问题：

(1) 保证信息系统安全的经典手段是“鉴别与认证”及“访问控制”，这种手段在经典的以及现代的安全理论中都是实行系统安全策略的最重要的手段。

传统的身份认证技术，并不能抵制脆弱性口令、字典攻击、特洛伊木马、网络嗅探器等攻击手段。对于访问控制，入侵者也可以利用系统漏洞、脆弱性程序等绕过访问控制，或提升用户权限；而且，访问控制等级和用户的使用效率成反比。

(2) 软件工程技术目前还没有达到 A2 级所要求的形式生成或证明一个安全系统的程度，Landwehr 等对计算机系统的安全漏洞进行分析表明，没有安全漏洞的计算机系统是不存在的[6]。目前，无论在理论上还是在实践中，试图彻底填补一个系统的安全漏洞都是不可能的，而要将所有已安装的带有安全漏洞的系统转换成安全系统需要相当长的时间。

(3) 目前还没有一种切实可行的办法解决合法用户在通过“身份鉴别”或“身份认证”后滥用特权的问题。

(4) 防火墙作为访问控制设备对于防范黑客有其明显的局限性。防火墙技术是内部网最重要的安全技术之一，但是，防火墙防外不防内，而据权威部门统计结果表明，网络上的安全攻击事件有 70%以上来自内部攻击；防火墙无法检测或拦截嵌入到普通流量中的恶意攻击代码，比如针

对 WEB 服务的 Code Red 蠕虫等。另外，防火墙难于管理和配置，易造成安全漏洞。

(5) 加密技术、访问控制和保护模型本身存在一定的安全缺陷[7]。

1.2.2　网络安全的 P²DR 模型与入侵检测

针对日益严重的网络安全问题和越来越突出的安全需求，“可适应网络安全模型”和“动态安全模型”应运而生，安全模型已经从以前的被动防御发展到了现在的主动防御，强调整个生命周期的防御和恢复，PDR 模型是最早体现这一思想的安全模型。所谓 PDR 模型是指基于防护(protection)、检测(detection)、响应(response)的安全模型。在 PDR 的基础上提出了 P²DR 模型，这里 P²DR 是 Policy(安全策略)、Protection(防护)、Detection(检测)、Response(响应)的缩写，P²DR 模型是一个动态的计算机系统安全理论模型，如图 1-3 所示。

图 1-3　P²DR 模型的体系框架

P²DR 模型为网络安全管理提供了一种新的方法。所有的安全问题都可以在统一的策略指导下，采取防护、检测、响应等不断循环的动态过程，它较传统静态安全方案有突破性的提高。P²DR 模型的具体内容如下：

(1) 安全策略。安全策略是 P²DR 模型的核心，所有的防护、检测、响应都是依据安全策略实施的，安全策略为安全管理提供管理方向和支持手段。在具体实施过程中，策略规定了系统所要达到的安全目标和为达到目标所采取的各种具体安全措施及其实施强度等。

(2) 防护。一般而言，防护是安全的第一步，具体包括制定安全规则、进行系统安全配置、采取各种安全措施等。通过修复系统漏洞、正确设计开发和安装系统来预防安全事件的发生；通过定期检查来发现存在的系统脆弱性；通过教育等手段，使用户和操作员正确使用系统，防止意

外；通过访问控制、监视等手段来防止恶意威胁。

(3) 检测。根据系统运行情况的变化，对系统安全状态进行实时的动态监控。检测是非常重要的一个环节，检测是动态响应和加强防护的依据，也是强制落实安全策略的有力工具，通过不断地检测、监控网络和系统，来发现新的威胁和弱点，通过循环反馈来及时做出有效的响应。

(4) 响应。发现了攻击企图或攻击之后，需要系统及时做出响应，包括：通知管理员、记录入侵行为、采取措施以阻止入侵及恢复系统正常运行等。紧急响应在安全系统中占有重要的地位，是解决安全潜在性问题最有效的办法，从某种意义上讲，安全问题就是要解决紧急响应和异常处理问题。

计算机信息系统的安全是基于时间特性的，P^2DR 安全模型的特点就在于动态性和基于时间的特性。入侵检测技术就是 P^2DR 模型中的检测部件，它是 P^2DR 模型体现动态性的关键所在。P^2DR 模型中每个环节都是密切相关的，策略在模型中处于核心地位，决定系统要达到的安全目标及采取的防护手段，但从另一个角度看安全策略也是制定入侵检测系统中检测策略的一个重要来源，入侵检测要根据已知的安全策略，更好地配置系统的参数。当发现入侵行为时，入侵检测系统会通过响应模块改变系统的防护措施，改善系统的防护能力。

通过 P^2DR 模型可以看出，人们已经越来越意识到仅有防护是不够的，单纯地使用被动防护技术不足以保护计算机系统的安全。为了能及时地发现并报告系统和网络中的攻击迹象和异常事件，采取入侵检测这样的动态监测、预防或抵御系统入侵行为的主动安全防护机制，通过监控系统和网络的运行情况发现入侵或入侵企图，并采取相应的安全响应策略，从而有效地提高安全性能。显然，入侵检测是对传统计算机安全机制的一种有效补充，它的开发应用增大了网络与系统安全的保护纵深，因而成为目前信息安全技术热点之一。

1.2.3 入侵检测的作用

在网络安全体系中，入侵检测系统(Intrusion Detection System，简称 IDS)是唯一一个通过数据和行为模式判断其是否有效的系统。如图

1-4所示，防火墙就像一道门，可以阻止一类人的进入，但无法阻止同一类人群中的破坏分子，也不能阻止内部的破坏分子；访问控制可以不让低级权限的人做越权工作，但无法保证高级权限的人不做破坏工作，也无法阻止低级权限的人通过非法行为获取高级权限；漏洞扫描系统可以发现系统和网络存在的漏洞，但无法对系统进行实时扫描。

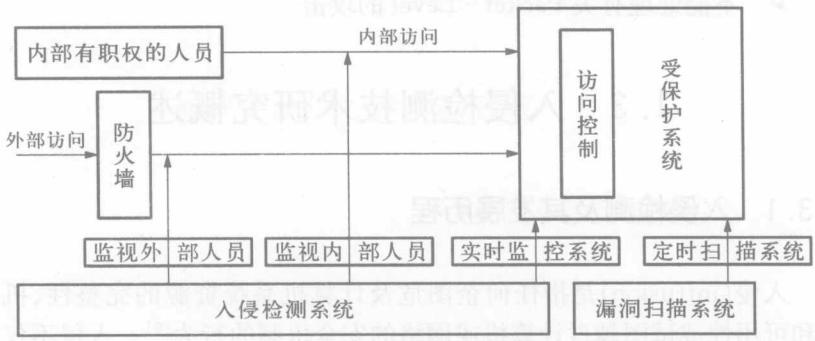

图1-4　入侵检测的作用

- IDS的主要作用和功能体现如下：
  - 监控、分析用户和系统的活动
  - 审计系统的配置和弱点
  - 评估关键系统和数据文件的完整性
  - 识别攻击的活动模式
  - 对异常活动进行统计分析
  - 对操作系统进行审计跟踪管理，识别违反政策的用户活动
- 入侵检测的优点如下：
  - 提高计算机信息安全构造的其他部分的完整性
  - 提高系统的监控性能
  - 从入口点到出口点跟踪用户的活动
  - 识别和汇报数据文件的变化
  - 侦测系统配置错误并纠正它们
  - 识别特殊攻击类型，并向管理人员发出警报，进行防御
- 入侵检测的缺点如下：
  - 不能弥补差的认证机制

- 如果没有人的干预，不能管理攻击调查
- 不能知道安全策略的内容
- 不能弥补网络协议上的弱点
- 不能弥补系统提供质量或完整性的问题
- 不能分析一个堵塞的网络
- 不能处理有关 Packet - Level 的攻击

1.3 入侵检测技术研究概述

1.3.1 入侵检测及其发展历程

入侵(intrusion)是指任何企图危及计算机系统资源的完整性、机密性和可用性或试图越过计算机或网络的安全机制的行为[8]。入侵不仅包括发起攻击的人(如恶意的黑客)取得超出合法范围的系统控制权，也包括收集漏洞信息，造成拒绝服务等对计算机系统造成危害的行为。入侵可能由通过互联网访问系统的攻击者发起，或由系统的某些授权用户发起，用户在错误地使用授予他们的特权时，也将造成对系统的入侵。入侵检测顾名思义，便是对入侵行为的发觉，它是为保证计算机系统的安全而设计与配置的一种能够及时发现并报告系统中未授权或异常现象的技术，是一种用于检测计算机网络中违反安全策略行为的技术。入侵检测系统通过对计算机网络和主机系统中的关键信息进行实时分析，从而判断出非法用户入侵和合法用户滥用资源的行为，并做出适当响应。它在传统的网络安全技术的基础上，实现了检测与响应，起着主动防御，从而使得对网络安全事故的处理，由原来的事后发现发展到了事前报警、自动响应，并可以为追究入侵者的法律责任提供有效证据。因此，入侵检测技术的出现使网络安全领域的研究进入了一个新的阶段。

入侵检测系统不同于其他安全产品，要求具有更多的智能，必须可以将得到的数据进行分析，得出有用的结果，并采取适当的对抗措施。对入侵检测的研究最早可追溯到 20 世纪 80 年代，但受到重视和快速发展还是在 Internet 兴起之后。文献[9 - 20]中对入侵检测技术的发展状况从