基于数据挖掘的入侵检测模型及其应用①

尤春梅 毛国君 鲁 杰 (北京市多媒体与智能软件重点实验室、北京工业大学计算机学院 100022)

摘要:入侵检测作为一种新一代的信息安全技术,需要在精度和效率上得到提高。引入数据挖掘等智能手段是提升入侵检测技术性能的关键。本文就数据挖掘技术在入侵检测系统中的研究和应用进行归纳和探索,并提出了基于聚类分析技术的异常入侵检测模型的构建流程和基本构件。

关键词:数据挖掘 聚类分析 信息安全 入侵检测

1 入侵检测系统研究现状与发展

入侵检测系统(IDS)从最初的实验室研究课题到目前的商业产品,已经具有 20 多年的发展历史。按照检测方法的不同,入侵检测技术可分为异常检测(Anomaly Detection)与误用检测(Misuse Detection)两种基本方法。异常检测是基于对正常操作的行为特征提取和总结的,即当用户活动与正常行为有重大偏离时即被认为是入侵。误用检测是基于对异常操作行为的特征提取和总结的,即当监测的用户或系统行为与异常操作行为特征接近时,被怀疑为入侵者。按照审计数据来源,入侵检测系统可以分为基于主机的入侵检测系统(Host - based IDS)和基于网络的入侵检测系统(Network-based IDS)两个基本类别。基于主机的入侵检测系统审计来源于系统运行所在的主机,保护的目标也是系统运行所在的主机。基于网络的入侵检测系统的审计数据来源于网络传输的数据包等,保护的目标是网络的运行环境。

入侵检测提出和发展具有很强的应用驱动。1980 年,Anderson 在"Computer Security Threat Monitoring and Surveillance"报告[1]中首先详细阐述了入侵检测的概念,提出了利用审计数据监视入侵活动的思想。1987 年,Denning 的"An Intrusion Detection Model"论文[2]可以称为入侵检测模型研究的开创性工作。1988 年,Denning 和 Teresal 公布了他们的入侵检测专家系统(Intrusion Detection Expert System,简称 IDES)[3],堪称 IDS 原型研制的典范。1990 年,加州大学戴维斯分校的 Heberlein 等人开发出了 NSM(Network Security Monitor)系统[4],将网络流作为审计数据来源,在网络环境下实现监控异种主机和网络设备的入侵检测。二十世纪 90 年代研究重点转到分布式入侵检测系统(DIDS)构架上。DIDS 既是一种系统体系结构的标志,同时也意味着多种技术的融合。分布式网络系统下的入侵可能是跨多个管理领域或操作系统的,数据的格式差异大、数据量膨胀,因此探讨大容量数据下多节点协同利用数据技术成为必需。围绕着协同工作和智能化等关键问题,DIDS 的实用或原型系统研制向着组件化方向发展,即面向入侵检测的过程研究高性能的组件。在此基础上设计一套协同入侵检测协议,入侵检测系统可以认为是由标准化的协同入侵检测组件组合在一起构成的。而在这些组件的设计中可以使用数据挖掘、神经网络、遗传算法、模糊与粗糙集、免疫原理等智能化技术来提高组件的效率和精度。二十世纪 90 年代中期后,众多组织和机构投入了将数据挖掘技术应用于入侵检测数据分析的研究,成为一个新的研究热点。

2 数据挖掘技术在入侵检测数据分析中的研究与应用

针对入侵检测系统的适应性以及智能性等要求,作为数据分析技术的前沿领域,数据挖掘显示出了它在审计数据分析中的优势。这种发展趋势的代表是哥伦比亚大学的 Wenke Lee 研究组的相关工作。Wenke Lee 研究组分别从网络和主机两个方面进行了审计数据的挖掘处理,并利用数据挖掘的分类、聚类以及序列分析等技术,针对拒绝服务攻击(DoS)、远程攻击(R2L)、本地用户非法提升权限的攻击(U2R)和漏洞扫描等进行了系列研究。尽管工作仍在继续,但是这种开创性的工作是成功的,已经吸引众多的学者开始将数据挖掘技术应用到入侵检测系统中的研究工作,并且可以断定在一定时期内将是入侵检测研究的一个热点问题。

近年的研究表明,数据挖掘技术在信息安全中的应用具有广阔的前景。目前应用较多的数据挖掘方法有 4 类:关联分析,序列模型分析、分类分析和聚类分析。将数据挖掘技术应用到异常检测中,可以从大量的数据中智能化"浓缩"出

① 本文得到国家自然科学基金(No. 60173014)、北京市自然科学基金(No. 4022003)和北京市教委资金资助

一个或一组值来表示系统行为的概貌，并以此进行用户行为的异常分析和检测。关联规则挖掘可能是最早应用到异常检测系统的数据挖掘技术。文献提出了一种用预测序列对(Look Ahead Pairs)和邻近序列(Contiguous Sequences)建立系统正常序列模型的方法。文献提出了一种统计方法，确定那些经常出现在入侵数据中而很少出现在正常数据中的序列。文献中使用在正常数据集上训练出的决策树建立系统预测模型，而文献中使用神经网络方法建立这样的模型。文献提出了无指导的异常检测的概念，并使用机器学习的方法建立系统调用序列的概率分布模型，将小概率事件解释为入侵事件。文献则应用聚类及其孤立点(Outlier)分析方法建立系统模型，用以分析系统行为模式。文献介绍了著名的 MADAM/ID 系统，它是通过对系统特征属性的归纳，并应用一种快速的规则学习算法 RIPPER，自动概括出普遍化的检测规则。这些尝试性工作反映了数据挖掘技术在信息安全中应用的广泛性。

另一方面，数据挖掘技术本身的发展也为信息安全领域的应用提供新的手段。例如，流数据挖掘(Stream Data Mining)是面向于动态、连续和有序数据流中知识发现问题提出的。众所周知，数据挖掘技术的研究及其成功的算法都是基于数据库技术的，即是在静态的数据集中发现有价值的知识模式。但是对入侵检测应用系统而言，要求对异常数据的快速反映，这就提出建立了在动态流数据的归纳和分析技术的要求。必须解决诸如有限内存与无限流数据、流数据采集速率与挖掘效率以及动态增量更改模式等关键问题。流数据挖掘在国内外都处于理论积累阶段，借鉴一些初步的成果，可以针对入侵检测的应用特点进行深入研究。

我国在入侵检测方面的系统化研究起步较晚，整体上处于探索和理论积累阶段。近年来，一些大学或研究机构，如中科院、清华、复旦等，也开始探索数据挖掘等智能化技术在入侵检测的应用。我们认为，和国外研究相比，国内研究还需要在基础理论、模型与算法探索和原型系统研制等方面加大探索力度。

3　基于聚类分析技术的入侵检测模型

数据挖掘技术在入侵检测中有广阔的应用前景，涉及的内容很广泛，是近年来一个颇为关注的研究课题。聚类分析是其中最具代表性的工作。

聚类就是将数据对象分组成为多个类或簇，划分的原则是在同一个簇中的对象之间具有较高的相似度，而不同簇中的对象具有较大差别。作为一个数据挖掘中的一个功能，聚类分析能作为一个独立的工具来获得数据分布的情况，并且概括出每个簇的特点，或者集中注意力对特定的某些簇作进一步的分析。此外，聚类分析也可以作为其他分析算法(如关联规则、分类等)的预处理步骤，这些算法在生成的簇上进行处理。聚类分析是一个活跃的研究领域，已经有大量的经典的和流行的算法涌现。

在传统的异常检测模型中，早期的统计分析都利用了参量的方法，描述用户或系统的行为模式特征。使用这种参量方法的前提条件时，我们所分析的数据满足某种特定的分布。在早期的 IDES 和 MIDAS 中，都假定用户模式满足高斯分布或正态分布。如果这种假设不成立，将会导致系统产生大量的错误报告。许多新的聚类方法得到探索并结合入侵检测问题进行有针对性的研究。例如，文献将不含入侵行为的 shell 命令序列聚类，并用各个类中心代表系统行为模式；文献先将从网络收集到的源数据进行初始聚类，再将这些数据输送到神经网络模型中进一步处理；文献针对入侵检测问题研究 Y－MEANS 方法。此外，聚类分析还经常用于源数据的预处理。

从整体来看，入侵检测系统主要有数据收集装置(负责收集审计数据)、数据分析部件(负责分析数据和检测入侵)、控制器(做出警告和反应)三大部分，其中数据分析部件是入侵检测系统的核心，作用在于对数据进行有效的组织、整理和分析，发现攻击并根据分析的结果产生事件，传递给控制器做出反应。数据分析的方式多种多样，基于聚类分析的入侵检测系统通常建立一个检测模型，即从审计数据中抽象概括出的系统正常行为概貌或异常行为模式，以此作为检测入侵的依据，检测模型的优劣与入侵检测系统的性能直接相关。因此，检测模型又是数据分析部件的核心。在异常检测中，检测模型通常由用户行为的统计特征来组成。统计特征可以由多种算法来计算。许多数据挖掘的算法被应用于其中。

本文所述的模型是一个基于网络的、基于聚类的无指导异常检测模型。传统的异常检测模型需要大量纯净的、不含攻击行为的正常数据进行训练和学习，这样的数据不可能容易地从实际运行的系统环境中直接获得。人们往往需要搭建一个专门收集这些数据的环境，来模拟正常操作和各种入侵行为。这就使得其应用受到很大的限制。为了可以在未标记的、来自实际环境中的、混杂了正常数据和入侵数据的原始数据上进行训练和学习，就需要研究入侵检测数据的特点，开发或改造聚类算法及其参数以适应混合数据的处理。

一般地，基于聚类分析的入侵检测模型构建的流程如图 1 所示。

检测模型的构建要经过历史审计数据的收集和存储、数据的预处理、聚类、类的标记(将生成的类标记为正常或异常)等几个阶段。每个阶段的工作重点不同，经过反复地测

试、比对检验以及算法修正后才可能达到理想的结果。

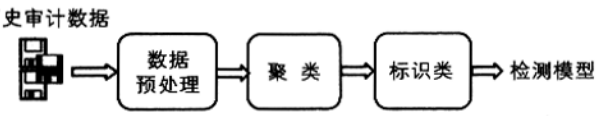

图 1　检测模型的构建流程

(1) 网络审计数据的收集：在以太网的一个冲突域中，通信基于广播方式，所以网络接口都可以接收到在此域的物理介质上所传输的所有数据。网络接口(网卡)一般有 4 种接收模式：广播模式、组播模式、直接模式、混杂模式。正常情况下，网卡的配置是同时支持前 3 种模式，只响应这样的两种数据帧：与自己的 MAC 地址相匹配的数据帧、发向所有机器的广播数据帧。若将网卡设置成混杂模式，则网卡即可"抓取"到所有流经过本接口的数据包。网络嗅探器(Sniffer)就是基于这样的工作原理进行"抓包"的工具，可用它完成源数据收集工作。有多种成熟的 Sniffer 产品可用，如 Tcpdump、Sniffem 等。

(2) 数据的预处理：此模块是核心模块之一，主要工作包括两个方面：由原始数据包括合成 TCP 连接纪录，构建特征属性空间。利用 Sniffer 抓取的是许多单个数据包的头信息和部分数据信息，数据量巨大。为减少分析处理的数据量，可以将归属于同一次 TCP 连接的数据包组合成连接纪录的形式。由于网络事件通常在时间上具有很强的相关性，尤其对于探测攻击(Portscan，Ping－sweep 等)及拒绝服务攻击(SYN－Flood，Teardrop 等)来说更是如此，因此，考虑在检测数据中加入基于时间的统计特性，这相当于在更抽象的层次上观察数据。可以采用的方法是使用时间窗的概念，针对每一条连接纪录，统计出在指定时间窗内与当前连接记录在属性上存在某种联系的连接纪录。包括以下两种统计方式：

- 监查在一个时间窗口内目标地址是某台主机的记录；
- 监察在一个时间窗口内目标端口是某服务端口的记录。

特征属性的选取也是非常重要的。选取的特征属性应能完整全面的反映系统特征，包括单个 TCP 连接的基本特征、重要的数据内容特征、连接的统计特征等[3]。

(3) 聚类算法开发：此模块是整个入侵检测模型的最关键的模块，主要工作是训练数据的规范化和聚类算法的选择、开发、运行和改进。为了将数据进一步转换成适合于聚类分析的形式，需要将数据进一步规范化，其中包括数字型属性的处理和离散型属性的处理。一个合理的规范化方法是 z－score 法。在进行算法开发时，可以根据入侵审计数据特殊的数据分布特征选择适合的现有聚类算法供参考，并利用划分、索引、组合等技术提高算法的性能。

(4) 类的标识：即将聚类过程中产生的各个类加标记，或者正常类(normal)，或者为异常类(anomaly)。由于检测模型的训练数据是混杂了无标识的正常数据和入侵数据，聚类过程结束后，相近模式的正常数据分别聚成簇，相似类型的攻击数据也分别聚成簇，需要分辨出那些是正常簇，那些是异常簇。由于正常数据远远多于异常数据，可根据簇中所含数据的多少，并结合相关的孤立点检测技术来标识它们的性质。

(5) 模型的测试：将测试数据与检测模型中的各个类相对比，与它最相近的类的标识即为此数据的标识。如果测试数据远离模型中所有的类，也将其标识为异常数据。从测试的结果可以评价模型的性能，若性能不理想，需要分析结果产生的原因，对前面的相应模块进行改进。

总之，传统的入侵检测是一个基于经验积累或简单模式匹配方法的，缺乏理论完备性和智能分析手段，因此这样的系统概括性差，只能发现模式规定的、已知的入侵行为，难以发现新的入侵行为。数据挖掘作为一种致力于从大数据集中发现知识的智能化手段，可以从海量安全审计数据中自动提取出尽可能多的隐藏安全信息，近几年在入侵检测系统研究中得到应用，成为信息安全中一个热点问题。

参考文献

1 Anderson J. Computer security threat monitoring and surveillance. Fort Washington, PA: James P. Anderson Co., 1980.
2 Denning D E. An intrusion detection model. IEEE Transactions on Software Engineering, SE－13: 222－232, 1987.
3 Teresal L, Jagganthan R, et al. IDES: The enhanced prototype, a real time intrusion detection system. MenloPark, CA: SRI International, Computer Science Lab, 1988.
4 Heberlein L T. A network security monitor. Proceeding of the IEEE Symposium on Research in Security and Privacy, Oakland, CA: IEEE: 296－304, 1990.