基于数据挖掘技术的网络入侵分析与检测

田春平[1]，刘　芸[2]

（1. 中国电信股份有限公司云南分公司，昆明 650000；2. 中国电信股份有限公司昆明分公司，昆明 650000）

摘要：入侵检测系统（Intrusion Detection System，简称 IDS）作为防火墙的补充，分为基于主机的入侵检测系统（Host-based Intrusion Detection System，简称 HIDS）和基于网络的入侵检测系统（Network Intrusion Detection System，简称 NIDS），至今被广泛的应用在我们的计算机上。但是传统的入侵检测系统存在自适应差、误报漏报率高、响应不及时等缺陷。近年来，“数据挖掘（Data mining）”出现在人们面前，它通过特殊的算法从海量的数据中搜寻有用的的信息，这些有用的信息可以用于预测发展趋势、关联不同事物、决策的支持等等。本文将数据挖掘技术和基于网络的入侵检测技术有机的结合起来，建立一个具有自适应性的网络入侵检测系统模型，解决传统入侵检测系统适应性差、漏报率高等缺陷。

关键词：数据挖掘；入侵检测；分类算法；模型

doi：10.3969/J.ISSN.1672-7274.2019.05.046

中图分类号：TP393.08　　**文献标示码：**A　　**文章编码：**1672-7274（2019）05-0065-03

针对传统的入侵检测系统只能检测处理已知的危险，漏报率高，需要网络与信息安全员手动更新规则库等缺点。通过引入数据挖掘技术，利用数据挖掘的特殊算法预测未知的攻击或危险，自动更新规则库。此方法可以解决传统入侵检测系统的短板，使系统及平台的受保护系数大大增高。

1　数据挖掘概述

1.1 数据挖掘的基本概念

数据挖掘产生于金融领域，通过对海量数据的分析和计算，得出有价值的信息，为决策提供数据依据。数据挖掘有着鲜明的特点，它对数据的处理是没有具体的假设，通过对大量数据的分析，发现其中那些表面上看不出来的信息，如“尿布和啤酒”能发现事件内在的隐藏联系，这些是人们发现不了了；或者能预测怀孕，通过分析计算，去了解顾客的生理特点，做出相应的政策，拉动经济增长；又或者通过分析事物内在联系，预测传染病的发生，提前预防等。这些都不是依靠人的经验或者直觉可以得到的知识，挖掘所得到的信息越是超出人的认识，其价值可能越大。传统的数据分析只能通过小量的数据分析得出结论，其信息的广度和准确度都不如数据挖掘。

1.2 数据挖掘技术的基本功能

基本功能主要体现在分类与回归、聚类分析、关联规则、时序模式和异常检测等。这五个基本功能主要解决的问题见表1。

表1 数据挖掘解决的实际问题

类型	主要解决的问题
分类与回归	将用户分等级，通过信用高低，给予不同服务。 预测客户接下来是否会继续购买。 预测对此产品感兴趣的人的特点。 预测治疗方案，为医生提供参考。 预测一天的人均销售额。 预测将会升级服务的客户。 预测房地产开发中的风险
聚类分析	通过对于一些事物，用其特定的表现来描述总结。 预测谁是银行信用卡的黄金用户。 预测不同客户会对同一类服务有喜好。 对不同路段聚类分析，得出店铺在哪里销售量会最高
关联规则	银行方面：通过顾客消费记录，分析其喜好。 某类型的设备故障是对应的某些零件出问题。 同一个患者用不同的药，康复情况不同。 某些商品具有先后关系，顾客买了第一个将会买另一个
时序模式	不同的月份，估计营业额。 预测明天某地区的最高用电符合
异常检测	检测计算机系统的不正常操作行为

1.3 数据挖掘的系统结构

典型系统结构，见图1。

2　入侵检测系统的概述

2.1 入侵检测系统的简介

现在，Internet 环境的多样化，黑客的能力在变强，各个地方都存在隐患。网络信息的安全不是靠一个技术就可以解决的，需要结合多方面才能做到。入侵检测系统是通过运行之后的审计数据来进行监控，这使得入侵检测系统的配置简单和可移植性高，几乎任何计算机系统都能使用。它的优势还在于有实时和动态监控等检测系统状态的方式。它的发展，提高了系统的安全性。随着技术的更新和升级，入侵者的知识能力也在逐渐提高，系统漏洞不断的被发现，入侵检测系统功能不够强大，所以人们对其的研究将会投入更多。

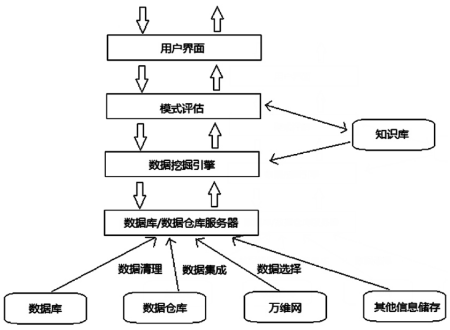

图1 典型结构

2.2 发展历史

早在1980年，对入侵检测的研究就已经开始，以下是具体的事例。

（1）1980年，美国的一篇文章，被称为入侵检测的开山之作。在这里，还提出了入侵检测的数据源可以从系统的审计数据中获取，但是当时并没有人们支持这个思想。

（2）1986年，为了检测数据库是否被不正常的操作，W.T.Tener 在 IBM 主机上用 COBOL 开发的 Discovery 系统，成为最早的基于主机的 IDS 雏形之一。

（3）1987年，乔治敦大学提出了 IDES（入侵检测专家系统），这是入侵检测技术第一次运用在计算机安全领域当中。它不需要考虑系统平台等问题，后来它作为一个通用框架被人们使用。

（4）1988年，因为遭受了“蠕虫”的攻击，网络瘫痪了近一周的时间，所以人们对入侵检测系统的研究更为热衷。

（5）1990年，通过对网络信息的分析，来确定是否有威胁安全的行为发生的想法被提出，成为入侵检测系统发展史上的一个分水岭。

（6）1994年，Mark Crosbie 和 Gene Spafford 建议使用自治代理（autonomous agents）以便提高 IDS 的可伸缩性、可维护性、效率和容错性，该理念非常符合正在进行的计算机科学其他领域（如软件代理，software agent）的研究。

（7）1995年开发了 IDES 完善后的版本—NIDES（Next-Generation Intrusion Detection System）可以检测多个主机上的入侵。

（8）1996年可能跨过多个管理领域的GrIDS（Graph-based Intrusion Detection System）问世，解决当代绝大多数入侵检测系统伸缩性不足的问题。

（9）1998年Ross Anderson和Abida Khattak将信息检索技术引进到入侵检测。

2.3 入侵检测系统的构成

基本组成部件如下：

（1）事件产生器：其功能是用于原信息的收集，它可以通过收集网络信息、系统的信息等。将这些表示为事件，是检测的开端，为接下来的部件传递此信息

（2）事件分析器：检测的核心，它接受了产生器给的信息之后，对信息做出分析操作，得出此信息是否安全。如果不安全则转换为信息传下去，如果安全则不用提示。

（3）事件数据库：其功能是存储各种信息，它同时接受以上两个部件发来的信息。一个是发出的事件信息，另一个是在对事件计算后产生的信息。

（4）响应单元：相应单元相当于事件处理模块，它将事件分析器发出的结果信息进行处理，它可以单纯的只是发出警报提示，等待人工处理，它也可以做出自动断开网络、处理文件等行为。

以上四个组件相互连接，见图2。

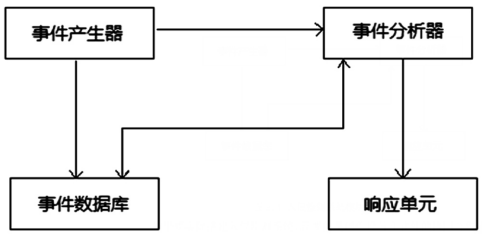

图2 入侵检测系统模块连接

如果要实际搭建入侵监测系统，需要的是硬件设备和软件的协同合作才能才能完成。事件产生器进行数据的收集这一块的功能需要通过硬件设施去实现，还有相应单元的相应报警、切断网络等功能也需要硬件的支持。

2.4 入侵检测系统模型

早期入侵检测模型见图3。

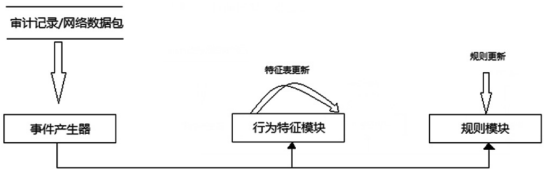

图3 早期入侵检测模型

改进后的入侵检测模型见图4。

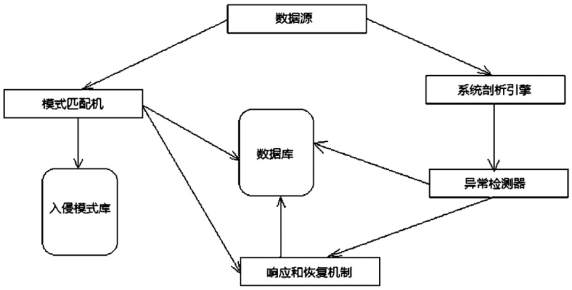

图4 改进的入侵检测模型

2.5 研究入侵检测系统的必要性

由于Internet的优势，计算机的功能愈加强大，加上网上交易的兴起，使得它们的身影存在各处。与此同时，不法分子因为个人利益而运用黑客技术去侵犯他人的个人信息及个人利益的事件也无处不在。他们利用系统的漏洞来入侵他人的计算机，人们为了抵御这些入侵做了许多的措施，如登录需要密码、验证码、密码学的加密解密等，但是这些并不够。

操作系统的隐患，见表2。

表2 系统存在的问题

	存在的问题
弱口令	弱口令很容易被盗取及爆破，入侵者就可以随意获取或者修改信息，它可以使得入侵者可以冒充主人，而且不会被拦截
静态安全措施不足	这个只是一个死的标准，使用者的行为多种多样，没有特定的界限来判断，其结果就是要么还有隐患，要么就是影响使用
系统软件	软件架构及编码缺陷被不断披露
修补系统软件缺陷并不常令人满意	系统软件涉及的方面很多，想要达到完美很难

由于计算机系统的如此多的漏洞无法在短时间解决，所以对于入侵检测的研究是势在必行的。

3 基于数据挖掘的网络入侵检测系统

随着科技的发展，网络已经渗透到生活的各个方面，而另一方面，大规模的网络入侵事件的发生，以至于网络安全已经成为焦点所在，而入侵检测系统就是其中的一个热门技术。传统的入侵检测系统只能检测被发现的，并存在于规则库里的，而对未知的安全隐患的检测极其低下。规则库的更行也是通过网络与信息安全员的人工操作，不仅费时费力，造价昂贵，而且效率低下，更新缓慢。随着黑客的能力的变强，传统的系统已经不够用了，需要提高其安全性能。

本章通过对以上内容的归纳总结，阐明数据挖掘的能为我们干什么事情，阐明网络入侵检测少些什么。并找出两者的交织区域。建立一个基于数据挖掘的网络入侵检测系统的模型，可以解析未知信息，并自动升级库。

3.1 结合过程

将两者结合主要的思想有连两个。一是通过分析，挖掘出网络入侵存在的一些未知的知识，如入侵行为的某些属性通过计算得出的值存在某些特殊性、入侵行为和某些数据有联系等等，直接识别不安全信息。另一个是通过分析，将用户安全的信息存入库，识别安全行为，间接识别非安全行为。本文构建的模型在这两个方向上都能使用。

在建立模型前，需要先确定数据挖掘和网络入侵检测系统的交汇点。数据挖掘技术通过对大量数据的分析和处理，最后得出某些结果，将这些结果处理成某些直观的表现方式，将其存入数据库；而网络入侵检测系统的重点在于规则库的内容，通过对比规则库确定当前数据的安全性。所以我们确定，两者的交汇点在于规则库，即将挖掘得出的结果存入库中。

使用跨行业数据挖掘的标准过程进行数据挖掘建模步骤：

第一步业务理解：我们首先要明白我们需要进行分析的数据是什么？我们需要分析的是网络数据。将带有隐患的网络信息分类，并且对库自动更新。

第二步数据理解：对于本文的数据来源于捕获的网络数据包，我们可以通过设置网卡为“混杂模式（Promiscuous Mode）”，这个模式使我们收集到和我们同一网段里的全部信息，也可以从网上下载数据包集。

第三步数据预处理：信息处理则是将数据包的包头信息提取出来作为属性，用于之后的数据挖掘。

因为我们数据挖掘的目的是为了构造入侵检测系统的自主学习，这个属于机器学习，而它的计算方式，必须要提供用于计算的数据。对于经过去噪声、去冗余、计算缺失值之后的数据，在这里是数据包，包含可很多信息，哪些可以用于计算呢？需要对这些经过初步筛选的数据进行怎么样的处理呢？

计算机文件，文件属性可以表示这个文件的特征，如：文件格式、文件大小、创建时间、修改时间等，这个属性能使我们不必要打开文件浏览文件内容，而对这个文件有一定的了解。如同文件一样，如果我们想通过解开数据包读取其内容的方式来确定

其安全性，这样的工作量太大，而且对系统的运行会是很大的影响，这是一个不可能实现的事。所以，我们准备通过网络数据包的文件属性入手，将其作为数据挖掘计算的数据。我们可以通过工具对网络数据包进行处理，将网络数据包的包头的信息提取出来，这些属性有：起始时间、源 IP 地址、目的 IP、传送字节等等。

第四步建立模型：系统的安全信息的数量比有隐患的信息数量是多得多的，如果收集系统的安全信息，这是一个工作量非常庞大的工作。相对来说，误用检测需要收集的数据只是有入侵嫌疑的数据，这个数据量的少了很多，故本文选用误用检测。建立的初步模型如图5，首先收集数据，如果是基于主机的，原始审计数据就从系统里收集，如日志文件等，如果是基于网络的，则可从网络收集，如用 TCPDUMP 工具进行收集；其次将数据的格式处理成 ASCII，并去完成其余预处理工作；然后对得到的信息进行筛选，创建模式，将选出的特征交给下一部分，并对入侵检测模型反馈的结果进行重新筛选特征，创建新的模式，再交给入侵检测模型，重复这个过程，到选出符合要求的特征为止；最后入侵检测模型将接受到的特征数据通过数据挖掘算法进行计算得出的结果，在对此结果进行评估，反馈给上一层，若不符合要求，则等上一层发来的新特征数据再运算，再对结果进行评估和反馈，重复操作，直到得到符合要求的模式。

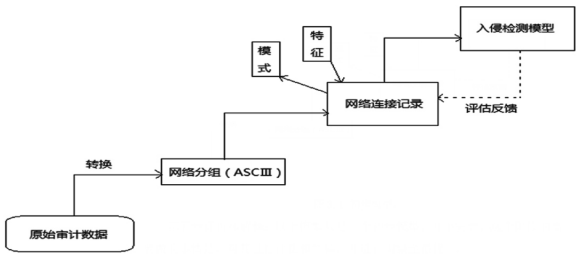

图5 初步模型

第五步评价和解释：以上模型只是一个初步模型，并不完善。这个阶段的需要做的事情是，对其进行评估和解释，并进行调整至最优。

3.2 基于数据挖掘的网络入侵检测模型

对于以上说到的初步模型，我们主要对其进行了优化，并进行了功能模块的完善，如图6所示，这是一个完整的基于数据挖掘的网络入侵检测系统模型。

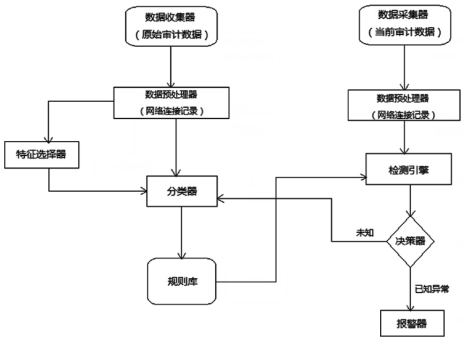

图6 完善之后的模型

在第一次运行前，左边模块要先运行，需要先对库做一个录入数据操作。我们可以从同一网段收集信息，但此办法的信息可能存在类型上比较单一、数据量不够等问题导致收集到的数据不足够支持接下来的数据挖掘运算，所以我们可以考虑从网络上下载前人已经获取好的较为完整的数据集作为我们的原始审计数据。当需要的数据准备好了之后，下一步进行预处理操作，在这里我们应该将收集的数据包进行获取其包头的信息，这些数据将交给分类器进行分类。但是这些处理好的数据并不是全都能用于数据挖掘的算法，甚至会影响结果，所以一般情况下需要对预处理完的数据进行筛选，经过多次的评价反馈，选出符合要求的特征信息，再将其交给分类器。分类器中的算法可以有很多种，如聚类、分类、关联等，具体的某些算法在第五章进行叙述。在第一次运行的时候，分类器需要对原始的审计数据进行运算，抽象出其中的规则（可以是系统正常运行的规则，也可以是具有安全隐患的规则，因为系统正常运行的情况的数据量太大，所以本文采用的是抽象出具有安全隐患的规则），将其存入规则库，之后运行时，只需要在决策器发出未知信息的时候，分类器再对未知的信息进行运算，并更新规则库。若未知的信息量大，每次都要去调用分类器，则分类时时刻刻都在运算，这样势必会影响系统地运行效率，我们可以将分类器设置为定时的运算，比如30秒，在这30秒之间到达的位置信息，先将其存放在缓冲区，当到30秒这个时间结点的时候，分类器再启动对这些未知信息进行运算，这样就可以解决拖慢系统运行速度的问题。规则库如上所述，分类器第一次运算得出规则作为初始规则存入规则库，并向入侵检测模块的检测引擎提供规则进行对比检测。在分类器不断的对未知数据运算的时候，不断的调整和更新规则库，使规则库越来越全面。规则库是入侵检测系统的核心，解决了规则库的更新问题，也就解决的传统入侵检测系统自适应性差，误报漏报率高（我们如果将系统正常运行的情况作为规则，这叫异常检测，这样的漏报率低，但是误报率高；如果我们将具有安全隐患的情况作为规则，则相反，这也叫做误用检测）的问题。

右边模块的运行获取的是当前的网络信息，进过一系列的处理后（和左边模块相同），将其交给下一个部件。检测引擎拿到上面传来的信息后，对其进行运算，得出来的结果去对比库里的信息。这是一个比较的过程，可以利用相似度比较的算法，得出的结果再交给决策器。决策器得到结果之后对其进行判断，若匹配失败，则认为这是一个安全度未知的数据，则将其交给数据挖掘模块的分类器进行运算；若匹配成功，就看作这是一个具有安全隐患的数据，则将其交给报警器，发出警报或自动处理。此模型针对误用检测，稍加修改也可以用于异常检测。

4 结束语

本文开始通过对数据挖掘技术的原理、发展、模型等的学习，重点要说明数据挖掘的强大功能，接着对入侵检测系统进行较全面的学习，然后综合了两者，将其结合，并构建了基于数据挖掘的网络入侵检测系统的模型，还对数据挖掘的算法进行了学习和初步的改进。

构建的基于数据挖掘的网络如入侵检测系统模型，是针对传统入侵检测系统无法自动更新规则库，对未知的情况无能为力，以至于传统入侵监测系统存在自适应性低，误报漏报率高等缺陷，结合了数据挖掘的强大能力，实现一种机器具有自我学习的能力，可以自主分析未知行为的安全性，并自动更新规则库，这就解决了问题。本文的设计都属于理论阶段，将来将进一步对其进行实验证明，并对其进行改进，如改进其可以检测网络数据包包头以外的信息。

参考文献

[1] 梁亚声等 . 数据挖掘原理、算法与应用 [M]. 北京：机械工业出版社，2014：3-22.
[2] 罗守山 . 入侵检测 [M]. 北京：北京邮电大学出版社，2003：12-40.
[3] 蒋建春，冯登国 . 网络入侵检测原理与技术 [M]. 北京：国防工业出版社，2001：1-27.
[4] Luo WJ，Zhang SH，et al.NIDS research advance based on artificial immunology[J].Journal of China University of Science and Technology，2002：520-541.
[5] 田生文 . 基于数据挖掘的网络入侵检测系统研究 [D]. 青岛：青岛大学，2005.