基于数据挖掘的入侵检测系统

An Intrusion Detection System Based on Data Mining

赵　悦　陈凌晖　(北京大学信息管理系 北京 100087,北京跟踪与通信研究所 北京 100094)

摘　要：建立完整的网络安全体系,入侵检测是主要的防御手段。本文针对传统入侵检测系统的局限性,提出一种利用数据挖掘技术,实现入侵检测的途径.本文提出的模型在网络的变化或升级时具有良好的扩展性,面对新型攻击模式具有较好的自适应性。

关键词：入侵检测 数据挖掘 关联规则 聚类分析

随着计算机技术与通信技术的发展,网络规模与复杂度不断增加,传统的,依靠人工分析的入侵检测系统已不能满足用户网络安全的需要。一方面,网络安全分析员不能高效地处理数量巨大的报警信息;另一方面,各种新的危险和攻击方式层出不穷,传统的入侵检测系统无法检测其未知的攻击行为,系统自适应性较差。

将数据挖掘技术应用到入侵检测中后,将提升检测系统的智能化和自适应性,提高系统的效率和精度。

1　入侵检测系统(IDS)概述

入侵(Intrusion)是指任何试图危害资源的完整性、可信度和可获取性的动作[1]。入侵检测(Intrusion Detection):按 Webster 辞典定义,即发现或确定入侵行为存在或出现的动作。识别入侵企图和行为并对此作出反应的计算机软件硬件系统就是入侵检测系统。

入侵检测系统根据其检测数据来源分为两类[2]:

基于主机(Host – based)的入侵检测系统,从单个主机上提取系统数据(如审计记录等)作为入侵分析的数据源。

基于网络(Network – based)的入侵检测系统,而基于网络的入侵检测系统从网络上提取数据(如网络链路层的数据帧)作为入侵分析的数据源。

除上述两种基本类型外,目前一些系统结合使用两者方式,称为分布式入侵检测系统。

根据检测策略,入侵检测技术可分为误用检测和异常检测[3]:

误用检测模式需要对已知攻击行为和系统的安全漏洞进行分析和分类,并手工编制相应的规则和模式,监视程序通过匹配,来识别入侵。

异常检测模式依赖于系统开发人员的直觉和经验选择统计量,以建立程序运行的正常模式,通过比较当前的程序行为与建立的模式之间的差异,当发现两者之间又重大偏移时,即认为系统遭受入侵。

由于开发过程中的人为因素及难免的随意性,当前的 IDS 的可扩展性和自适应性还非常有限。

2　数据挖掘技术在入侵检测数据分析中的应用

数据挖掘的核心思想是发现模式,将数据挖掘技术应用在入侵检测系统中,可以从系统日志、网络流量等大量原始数据中发现有助于检测攻击的知识和规律,动态更新 IDS 系统的规则库,提高 IDS 异常行为的识别能力和未知模式攻击的检测能力。在入侵检测中所用到的数据挖掘方法主要有以下几种:

(1) 分类分析[4]。数据分类分为两个过程,首先选择一个训练数据集,其中包含标识训练样本事件所属类别的数据项,类别是已知的,利用分类规则、判定树等数据挖掘算法构建为每个类别作出准确的描述或建立分析模型或挖掘出分类规则。第二步,利用得到的模型或分类规则对收集的网络数据流进行分类。首先评估模型(分类规则)的预测准确率,对于每个测试样本,将已知的类标识与该样本的类预测标识进行比较,模型在给定测试集上的准确率是被模型正确分类的测试样本的百分比。如果模型的准确率可以被接受,就可用于对类标识未知数据的分类。

(2) 关联规则挖掘[5]。所谓关联规则,是指数据

对象之间的相互依赖关系,而任何两个数据对象间都可能存在着潜在的关联,因此在考察哪些关联确实具有代表性,真的很有作用,哪些关联只是假象或者毫无用处时,需要同时考虑两条独立的标准,即确信度(Confidence)和支持度(Support)。发现规则的任务就是从数据对象中发现那些确信度(Confidence)和支持度(Support)都大于给定值的强壮规则。

(3) 序列模式挖掘[6,7]。序列模式分析主要用于发现形如"在某段时间内,有数据特征 A 出现,然后出现了特征 B,而后 C 又出现,即序列 A→B→C 出现频度较高"之类知识。由于网络攻击与时间变量紧密相关,因此序列模式分析在关联分析基础上进一步分析攻击行为时间相关性。它主要挖掘安全事件之间先后关系,运用序列分析发现入侵行为的序列关系,从中提取入侵行为之间的时间序列特征。

(4) 聚类分析。用于描述和发现数据中以前未知的数据类型,其中样本数据中不包含类别变量,数据挖掘将具有共同趋势和模式的数据元组聚集为一类,使类内各元组相似程度最高,类与类之间差异最大。基于聚类分析的入侵检测算法基本思想主要源于入侵与正常模式上的不同以及正常行为数目的条件,因此能够将数据集划分为不同的类别,由此分辨出正常和异常行为来检测入侵。基于聚类的入侵检测是一种无监督的异常检测算法,通过对未标识数据进行训练来检测入侵,因此能发现新型的和未知的入侵类型[6,8]。数据挖掘中常用的聚类算法有 k – means、模糊聚类、遗传聚类等。

(5) 孤立点分析[4]。孤立点分析属于聚类分析的一种特例。孤立点(outlier)指不符合数据一般行为或模型的一些奇异数据,孤立点探测和分析的过程被称为孤立点挖掘。孤立点挖掘问题可以被看作两个子问题:(1)定义在给定的数据集合中什么样的数据可以被认为是不一致的;(2)找到一个有效的方法来挖掘这样的孤立点。孤立点检测的方法包括三类:统计学方法、基于距离的方法和基于偏移的方法。

3 基于数据挖掘技术的入侵检测原型系统

3.1 原型系统结构

原型系统主要由探测器、数据接收模块、数据库模块、特征提取模块、数据挖掘模块、模式规则库、检测模块等七大部分组成,如图 1 所示。每个模块的具体功能如下:

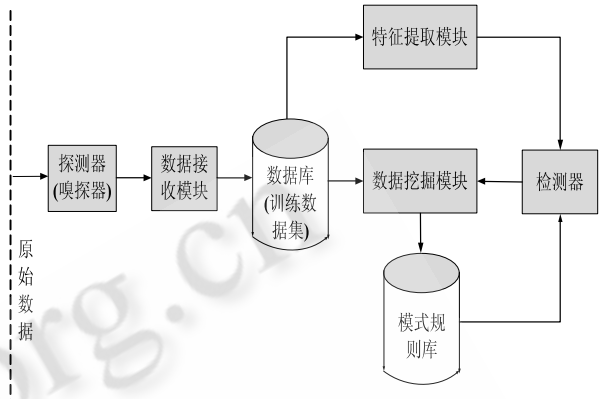

图 1　原型系统结构框架

(1) 探测器接收来自网络的原始数据,并将数据转换为标准的 ASCII 格式。探测器就是一个网络嗅探器(Sniffer),目前有多种成熟的 Sniffer 产品可用,如 Tcpdump、Sniffem 等。

(2) 数据接收器负责搜集来自传感器的数据并对数据进行预处理。经过预处理后的数据成为由多个连接特征组成的连接记录数据集合,特征包括连接持续的时间、连接使用的服务端口、连接的结束状态等[9]。然后利用分箱、聚类、回归等技术方法将这些数据集合转换成适合数据挖掘的形式。

(3) 数据库用于向经过预处理的数据提供半永久性存储。数据库中的数据即可作为训练数据集。

(4) 数据挖掘模块利用数据挖掘技术从数据库中提取有关行为特征和规则,建立入侵行为和正常行为轮廓,从而建立检测模型,存入模式规则库中。

(5) 模式规则库用于存放由数据挖掘模块产生的入侵行为和正常行为的规则。

(6) 特征提取模块采用数据挖掘技术对当前用户行为进行分析,从中提取出当前用户行为的由一系列属性组成的特征。

(7) 检测模块依据模式规则库中的入侵行为与正常行为的规则,对特征提取模块送来的当前用户行为特征进行比对分析。用户行为的分析结果可分为三类:用户正常行为、入侵行为和模块规则库不能判断的新情况。对于入侵行为系统发出报警信号;对于不能判断的新情况,检测模块将其输入到数据挖掘模块中进行二次挖掘。若发现入侵行为,发出报警信号,最后

将新挖掘出的正常行为规则与入侵规则存入模式规则库,进而完成对规则库的更新。

本文所述的原型系统是一个,基于网络的,具有自适应能力的无指导混合检测模型。具体来说,模型检测的数据来源是网络中捕获的数据包。模型挖掘出的正常行为模式和入侵行为模式全部存入模式规则库中,以供检测模块对比分析使用,因而模型采用的检测策略是综合异常检测和误用检测的混合式检。检测模块能够将规则库中没有的用户行为特征送到数据挖掘模块进行二次挖掘,并更新规则库,使模型能够识别出用户新的正常行为和新型的攻击行为,因此原型系统具有自适应能力。检测模块本原型系统与传统的基于数据挖掘的入侵检测系统不同,本模型中,数据挖掘模块使用的训练数据集是经过预处理的来自实际网络环境的原始数据。传统的基于数据挖掘的入侵检测系统则需要大量纯净的、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,本原型系统设计了一个基于关联规则、序列分析和聚类分析等算法的数据挖掘模块。

3.2 数据挖掘模块工作原理

正文内数据挖掘各种算法在应用中的侧重点和优势各不相同,在实际应用中,结合使用往往会有更好的效果。例如,关联规则算法可用于发现网络连接记录各属性之间的关系,序列分析算法能发现网络连接记录间的时序关系。使用关联规则和序列分析算法可以得到正常行为的模式。分类分析算法则可以挖掘出能识别出正常行为和入侵行为的规则。

原型系统中的数据挖掘模块综合使用了关联规则、序列分析、聚类分析和分类分析四种算法,模块的工作流程如下:

(1) 使用关联规则和序列分析算法处理训练数据集中的连接数据。首先通过关联规则算法得到连接记录特征属性的关联模式;由于网络攻击与时间变量紧密相关,因此使用序列分析在关联分析的基础之上进一步分析连接记录的时间相关性,从而得到序列模式,进而挖掘出正常行为模式。

(2) 用正常行为模式去过滤网络连接数据,从而得到纯度比较高的入侵数据,并重新建立训练数据集。在更新的训练数据集之上使用聚类算法进行挖掘。为了让训练集数据转换成适合于聚类分析的形式,需要对数据进行包括数字型属性和离散型属性的规范化处理。一个合理的规范化方法是 Z – score 法[10]。

(3) 为聚类过程中产生的各个类加标记,或者正常行为类,或者异常类。根据正常数据远远多于入侵数据的原理,依据类中所含数据的多少,并结合相关的孤立点分析算法来标识它们。

(4) 用分类分析算法同时对使用关联规则与序列分析算法得到的正常行为模式和通过聚类算法得到结果进行分类规则提取。

(5) 将上一步提取出的分类规则存入模式规则库中。

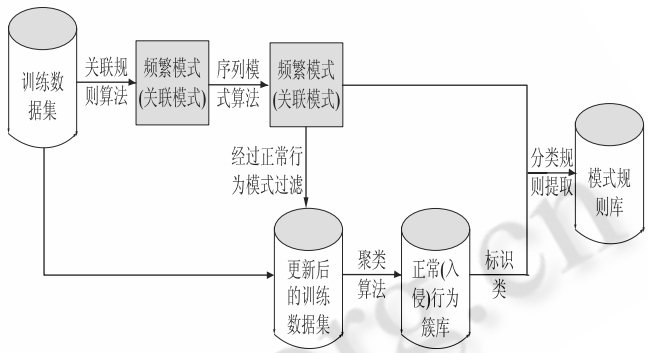

图 2　数据挖掘模块工作原理图

3.3 关联分析算法描述

常用的关联规则算法在对训练集进行关联分析时,容易得到伪肯定率较高的结果。伪肯定是指某一行为并非一次入侵,但是入侵检测系统却将其当成一次入侵攻。比如一台繁忙的服务器不断地接收大量的 SYN 包,但 IDS 却误将其当作一次"SYN 湮灭"攻击。针对这一问题本模型使用一种可有效降低未肯定率的关联规则算法。

首先挖掘出具有高支持度和置信度的规则集,具有稍低支持度和置信度的规则集以及推出入侵的规则集,然后比较第二者与第三者的相似度,得出能够有效降低伪肯定率的规则集,最后将其与具有高支持度和置信度的规则集进行合并得出正常行为规则集。

3.3.1 挖掘具有稍低支持度和置信度的规则集

输入:训练集合 T,初始支持度和置信度 s,c,阀值 α,b

输出:规则集合 P_1

While(T0) do

P_0 = miningSupportRule(T,s,c); //挖掘具有高支持度和置信度的规则

S = selectSupportRecord(P_0); //选出为规则 P_0 所支持的所有记录集

T = T − S; //删除为 P_0 规则所支持的记录集

P_1 = miningSupportRule(T,s − α,c − b); //挖掘具有稍低支持度和置信度的规则

return;

End

设|T| = n 中,函数 miningSupportRule(T,s,c),selectSupportRecord()的复杂度均为 O(n),则该算法的复杂度为 O(n)。挖掘推出否定类的规则集算法与该算法相似,在此不做讨论。

3.3.2 计算相似度的算法

输入:具有稍低支持度和置信度的规则集 P_1,推出否定类的规则集 N,相似度计算阀值 β

输出:规则集合 P′

```
Init P′ = P1; //初始时令 P′ = P1
For(i = 0i < |P1|; i + +)
For(k = 0; k < |N|; k + +){
 intsimilarity;
 similarity = Computer(Nk, P1i); //计算 N 中第 k 条规则和中第 i 条规则之间的相似度
 if(similarity > B){
  P' = P' − P1i; //减去具有高相似度的规则
  break;
  }
 }
 return P';
End
```

设|N| = n,|P_1| = m,显然该算法的复杂度为 O(mn)。

3.3.3 检测规则集的合并

由于 P_0 和 P'中没有重复的元素,故该算法只是将两集合中所有的元素加在一起形成一个新的集合,较为简单,在此不做讨论。

4 结束语

入侵检测系统是一个典型的数据处理系统。它通过对大量的数据进行分析,来判断被监控的系统是否受到了入侵攻击。其核心问题就是如何从已知数据中获取正常行为规则和入侵行为规则。基于此,本文提出一种基于数据挖掘技术的入侵检测原型系统,本模型的优势在于不需要专门搭建训练数据集,可使用实际网络环境中的原始数据进行训练,从而减少了开发成本,提高了系统的灵活性和实用性。此外,本模型尝试了一种新型的关联规则算法,可有效降低规则集的伪肯定率。由于模型中使用了较多的算法,如何提高数据挖掘模块的效率将是下一步工作的重点。

参考文献

1 R. Heady, G. Luger, A. Maccabe, M. Servilla. The archit ecture of a network level intrusion detection syetem. Technical report Computer Science Department University of New Mexico August 1990.

2 Stefan Axelsson. Intrusion Detection Systems: A Survey and Taxonomy. http://ww w. mnlab. cs. depaul. edu/seminar /spr2003/IDSSurvey. pdf, 14 March 2000.

3 Lee W, Salvatore J Stolfo. Adaptive Intrusion Detection:a Data Mining Approach, Artificial Intelligence Review, Kluwer Academic Publishers, 2000,14(6):533 − 567.

4 Lee W, Stolfo S J. Data Mining Approaches for Intrusion Detection. Proc. 7th USENIX Security Symposium,1998.

5 Agrawal, R. Srikant. Fast algorithm for mining association rules. In: Jorge, B. B, Matthias, J. Carlo, Z. eds. Proceedings of the 20th International Conference on Very Large Databases. Santiago: Morgan Kaufmann Publishers, Inc, 1994. 487 − 499.

6 J. Han, Micheline Kamber (著). 范明译. Data Mining:Concepts and Techniques. 北京:机械工业出版社, 2001. 20 − 116.

7 Box G E P, Jenkins G M. Time Series Analysis, Forecasting and Contronl, Holden − Day(2ed), 1976.

8 Portnoy L, Eskin E, Stolfo S J. Intrusion Detection with Unlabeled Data Using Clustering. Philadelphia: Proceedings of ACM CSS Workshop on Data Mining Applied to Security(DMSA 2001), ACM Press, 2001.

9 Lee W, Stolfo S, Mok K. Mining Audit Data to Build Intrusion Detection Models. The Fourth International Conference on Knowledge Discovery and Data Mining (KDD 98), New York, NY, 1998.

10 Taffler, R. J. The Use of the Z − Score Approach in Practice. Working Paper 95/1, 15th June 1995.