基于数据挖掘的入侵检测系统模型

程玉青,梅登华,陈龙飞
(华南理工大学 计算机科学与工程学院,广东 广州 510006)

摘 要:文中介绍了入侵检测系统的重要性、传统入侵检测技术的类型和局限性以及入侵检测系统中常用的数据挖掘技术,指出数据挖掘技术应用在入侵检测系统中的可行性和必要性。针对现有入侵检测系统存在的误报率和漏报率较高的问题,对数据挖掘技术应用于入侵检测系统进行了研究,提出一个基于数据挖掘技术的结合异常检测和误用检测的复合入侵检测系统模型,并对模型中的数据挖掘算法进行了探讨。实验表明,该模型能生成新规则,对新攻击具备一定的鉴别能力,能有效降低入侵检测系统的误报率和漏报率。

关键词:数据挖掘;入侵检测;复合引擎
中图分类号:TP393.08 **文献标识码**:A **文章编号**:1673-629X(2009)12-0123-04

A Model of Intrusion Detection System Based on Data Mining

CHENG Yu-qing, MEI Deng-hua, CHEN Long-fei
(School of Computer Science and Engineering, South China University of Technology, Guangzhou 510006, China)

Abstract: Introduce the importance of intrusion detection as well as types of traditional intrusion detection and their limitations, point out that data mining technology may overcome these limitations. To solve the problems of intrusion detection system, such as the high rate of serious distort and fail to report, an intelligent model of intrusion detection which uses misuse detection and anomaly detection based on data mining approach is proposed. Experiments show that the model can produce new rules, find new type of intrusion data and decrease the rate of false alarm and fail to report.

Key words: data mining; intrusion detection; compound engine

0 引 言

随着网络技术和网络规模的不断发展,网络在人们日常生活、社会、经济和军事中扮演着越来越重要的角色,人们对它的依赖程度日益加深;另一方面,由于黑客的入侵对国家安全、经济和社会生活造成了极大的威胁,使人们无法回避网络安全的问题。为了保护信息的安全和关键基础网络设施,入侵检测技术受到越来越多的重视,网络入侵检测系统已经成为网络安全架构的一部分。入侵是指违背访问目标的安全策略的行为。网络入侵检测系统(IDS)检测并向系统管理员报告这些行为,或者采取合法的措施来抵御它们。通过检测入侵,系统管理员知道哪些地方需要加强防范,哪些补丁还没有打,哪些用户需要加强安全。

入侵检测技术可被分为两大类[1]:误用检测(misuse detection)和异常检测(anomaly detection)。误用检测首先要建立攻击模式库,对系统审计数据和网络数据进行过滤,检查是否含有与攻击模式库匹配的入侵行为标识,主要采用模式匹配、专家系统和状态转移等技术。异常检测采取数学手段建立系统或用户行为的正常行为简档(normal profile),检查安全审计数据是否存在与之违背的异常模式,主要运用量化分析、统计分析、基于规则的检测和神经网络等技术,其关键在于确定一个恰当的异常阈值,否则将造成很多误报警或漏检。

针对误用检测中漏报率高和异常检测中误报率高的特点以及对报警进行数据挖掘时必须面对海量数据的特点,该文提出了将误用检测和异常检测结合的入侵检测框架。在该框架中,使用复合检测引擎清洗数据,并进行常规报警再通过挖掘引擎挖掘规则,在规则挖掘之前引入了k-means聚类算法,来聚类复合检测引擎清洗过的数据;然后使用Apriori算法挖掘聚类中

收稿日期:2009-03-31;**修回日期**:2009-06-17
基金项目:国家自然科学基金与中国民用航空总局联合资助项目(66776816)
作者简介:程玉青(1978-),男,山东茌平人,硕士研究生,研究方向为计算机网络安全;梅登华,博士后,副教授,研究方向为计算机网络安全及可信计算。

的频繁项并形成规则;最后,根据网络中正常数据和异常数据体现不同特征,对提取规则的正确性进行评价然后再发布到检测系统中。

1 数据挖掘在入侵检测中的应用

数据挖掘能分析原有的数据,做出归纳性的推理,从中挖掘出潜在的模式,预测出客户的行为。从原理上可以将数据挖掘的分析分为关联分析方法(Association)、序列模式分析(Sequential Pattern)等[2]。其中关联分析的目的就是挖掘出隐藏在数据间的知识,通过分析记录集合,推导出数据项之间的相关性,从而找出入侵者的各种入侵行为之间的相关性,主要的算法有 Apriori 算法和 AprioriTid 算法等。而序列模式分析与关联分析相似,但侧重于分析数据间的时间前后及因果关系。

1.1 基于关联规则分析的入侵检测

关联规则挖掘是数据挖掘最为广泛应用的技术,也是最早用于入侵检测的技术之一。关联规则(Association Rules)表示数据库 D 中一组对象之间某种关联关系的规则[3]。D 为具有 n 列属性的事务数据库,A 为数据库 D 的属性集合,事务 $T=\{t_1,t_2,\cdots,t_m\}$ 是 A 上的值集。设属性集 $X,Y\subseteq A$,如果事务数据库中有 s 的事务包含 $X\cup Y$,那么关联规则 A 的支持度(support)为 s;数据库中所有含属性 X 的数据集为 $D_x=\{T\mid T\in D\cap X\in T\}$,如果 D_x 中有 c 的事务包含属性集 Y,那么 $X\Rightarrow Y$ 关联规则的信任度(confidence)为 c。

入侵检测领域关联规则的应用是很有意义的,因为:

①整理后的挖掘数据可以方便地用数据库的表格形式表示,每一行表示一个连接记录,每一列表示记录中的一个字段(或某种属性特征),正是这种便捷的表示方法,使快速的检索和计算成为可能。

②用户相关性的行为在一定程度上具有重复出现的特征,正是这种频繁出现的行为特征可以用关联规则的形式表现。

③新产生规则可以很容易地不断加入到已有的规则集中。

1.2 基于聚类分析的入侵检测

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及正常行为数目应远大于入侵行为数目的条件,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。数据挖掘中常用的聚类算法有 k-means、模糊聚类、遗传聚类等。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。该方法不需要手工或其它的分类,也不需要进行训练,因此能发现新型的和未知的入侵类型。

1.3 基于序列模式分析的入侵检测

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,而后特征 C 又出现了,即序列 A→B→C 出现频度较高"之类高频序列信息[4]。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,从中提取出入侵行为之间的时间序列特征。序列模式分析一般不单独使用,它可用于入侵检测过程的某一步骤,从数据中挖掘用户序列模式,提取出可用于入侵检测的知识和模式,如对网络连接数据进行序列分析,正确提取出一些基于时间的统计属性,以便能构造出分类模型。序列模式分析与关联规则有相似之处,它对反复出现的序列检测率较高。

1.4 基于分类分析的入侵检测

入侵检测可以看作是一种数据分类问题。进行分类挖掘的入侵检测应首先选择一个训练数据集,对该训练集标记出正常或异常的数据,使用分类规则、决策树等方法从该数据集中提取出分类规则并构造出适合的分类器;然后用构造出的分类器对收集的网络实时数据流进行分类,将数据分为正常行为或某种入侵行为,以此判断出是否存在入侵行为。这一分类过程应该不断反复和评估,以期望能够得到最优化的分类器。分类分析一般是基于数据的特征属性,特征的选取对建立分类模型的准确性影响很大,因此单一使用分类思想进行入侵检测往往效果并不理想。为了提高检测准确率,需要将分类与序列分析、聚类分析等数据挖掘方法融合在一起来进行入侵检测。

2 检测模型框架

该模型框架由数据采集器、复合检测引擎和数据挖掘引擎三部分组成,如图 1 所示。

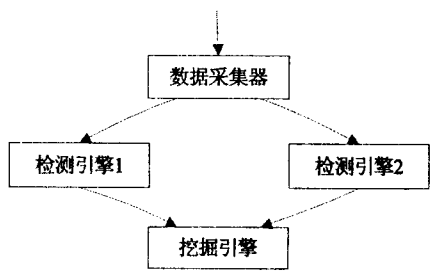

图 1 检测模型框架

数据采集器采集报警数据,是系统的数据源;复合检测引擎,结合了误用检测和异常检测这两种技术,有效消除了冗余数据;数据挖掘引擎是系统实现中的关

键,首先采用 k－means 聚类算法聚类复合检测引擎清洗过的数据,得到聚类数据,再使用 Apriori 算法挖掘聚类数据中的频繁项,然后将挖掘出的频繁项按照规则描述语言改写成为通用规则的形式,最后发布到检测引擎中。

2.1　复合检测引擎

在该模型中,采用一种复合结构,不同于以往的引擎,而是综合了传统的误用检测和异常检测的特点,其结构如图 2 所示。

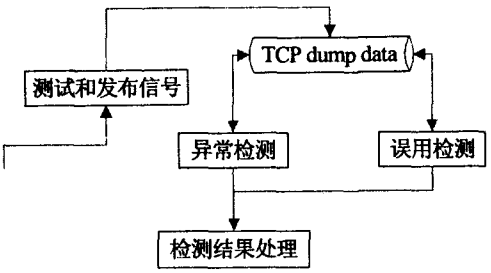

图 2　复合检测引擎

当数据采集部件采集的数据到达检测引擎后,检测引擎中的两个独立部件将会先后处理数据。误用检测引擎检测出明显的攻击数据,并告警;接着经过误用检测引擎处理后的数据再由异常检测引擎进行第二次过滤,检测出那些明显的正常数据,这样同时克服了误用检测漏报率高和异常检测误报率高的缺点。误用检测引擎和异常检测引擎分别处理数据采集部件采集到的数据,它们工作的先后顺序是随机的,也就是说,它们没有固定的先后顺序。同时,检测引擎还有一个任务,就是负责测试和发布挖掘引擎提取的规则。当测试部件接收到挖掘引擎提取的规则之后,首先将规则放入测试池,检测引擎在检测过程中,同时也需要对测试池中的规则进行匹配,如果测试池中的规则匹配成功,检测引擎不会告警,但会在规则的相应地方登记匹配的情况。测试部件每隔一段时间需要检查测试池中规则的匹配曲线,然后根据各匹配曲线,判断规则的可信度和规则所刻画的数据类型。并根据曲线将测试时间足够长的规则发布到对应的规则树,参与检测引擎的检测任务。

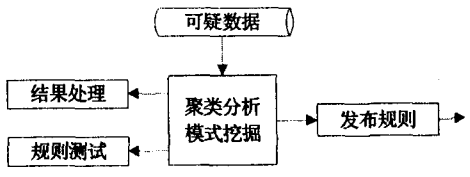

图 3　挖掘引擎

2.2　数据聚类

数据挖掘引擎结构如图 3 所示,首先使用 k－means 算法对数据进行聚类处理,以便于提取频繁项。k－means 算法接受输入量 k;然后将 n 个数据对象划分为 k 个聚类以便使得所获得的聚类满足:同一聚类中的对象相似度较高,而不同聚类中的对象相似度较小。聚类相似度是利用各聚类中对象的均值所获得一个“中心对象”(引力中心)来进行计算的。

k－means 算法步骤如下:

(1) 假设要聚成 k 个类,由人为决定 k 个类中心 $z_1(1), z_2(1), \cdots, z_k(1)$。

(2) 在第 k 次迭代中,样本集 $\{z\}$ 用如下方法分类:对所有的 $i = 1,2,\cdots,k, (i \neq j)$,若 $\| z - z_j(k) \| < \| z - z_i(k) \|$,则 $z \in s_j(k)$。

(3) 令由(2) 得到的新的 $s_j(k)$ 类中心为 $z_j(k+1)$,令 $J_j = \sum_{z \in s_j(k)} \| z - z_j(k+1) \|^2$ 最小。$j = 1,2,\cdots, k$,则 $z_j(k+1) = \frac{1}{N_j} \sum_{z \in s_j(k)} z$,$N_j$:$s_j(k)$ 中的样本数。

(4) 对所有的 $j = 1,2,\cdots,k$,若 $z_j(k+1) = z_j(k)$ 则终止,否则 goto 步骤(2)。

k－means 算法运算速度快,内存开销小,比较适合于大样本量的情况,但是聚类结果受初始凝聚点的影响很大,不同的初始点选择会导致截然不同的结果;并且当按最近邻归类时如果遇到到两个凝聚点距离相等的情况,不同的选择也会造成不同的结果,因此,k－means 具有因初始中心的不确定性而存在较大偏差的情况,需多次迭代,多次修正。

2.3　提取频繁项

在数据聚类之后,挖掘引擎需要提取聚类中的频繁项,在这里采用了较为通用的 Apriori 算法[5,6]。Apriori 算法是一种逐层搜索的迭代方法,k － 项集用于产生 $(k+1)$ － 项集。

算法步骤如下:

(1) 每个项都是候选 1 － 项集的集合 C_1 的成员。算法简单扫描事务数据库中的所有事务,对每个项的出现次数进行计数,这样就得到了候选 1 － 项集的集合 C_1。扫描 C_1,删除那些出现计数值小于阈值的项集,这样就得到 1 － 频繁项集的集合 L_1。

(2) 为找 L_k,通过 L_{k-1} 与自己进行连接产生候选 k － 项集的集合,该候选项集的集合就记作 C_k。

(3) 对 C_k 进行剪枝,从 C_k 中删除所有 $(k-1)$ － 子集不在 L_{k-1} 中的项集。

(4) 对事务数据库 D 进行扫描,将每个事务 t 与 C_k 中的候选项集 c 作比较,若 c 属于 t 则将 c 的计数值加 1(在扫描之前,初始值为 0)。扫描 C_k,删除那些出现计数值小于给定支持度的项集,这样就得到了 k － 频繁项集的集合 L_k。

(5) 循环执行步骤(2) 到步骤(4),直到 L_k 为空。

(6) 对 L_1 到 L_k 取并集即为最终的频繁集 L。

(7) 对于每个频繁项集 l 产生所有非空子集,然后对于其中的每个非空子集 s, 如果 sup port_count(l)/sup port_count(s)≥min_conf,则输出规则,其中,min_conf 是最小置信度阈值。

Apriori 算法是一种以概率为基础的具有影响的挖掘布尔型关联规则频繁项集的算法,它已被广泛用于商业决策、社会科学、科学数据处理等数据挖掘领域。

2.4 规则的测试与发布

挖掘出的规则需要部署到检测引擎中测试,经过测试才可以发布到检测引擎作为检测的判据。规则测试的主要依据是根据网络中数据的时间特性,一般情况下,正常的数据是连续的,同时存在一定的自相似性[7,8]。常数据具有阵发特性,通常大量的数据只在某一段时间出现。

根据这一特点,将匹配测试规则的数据按时间分布计,然后分析数据的时间特性,就可以轻易地判断规则的类别,然后再根据规则的类别将规则发布到误用检测引擎或检测引擎的规则库,从而实现规则的自动提取和发布。

3 实验结果

为评价该模型性能,以开源 snort 为基础,分离 alert 规则和 pass 规则,构建一个简单的复合检测引擎。基于 DARPA1999 数据集进行测试,四周测试结果如图 4 和图 5 所示。

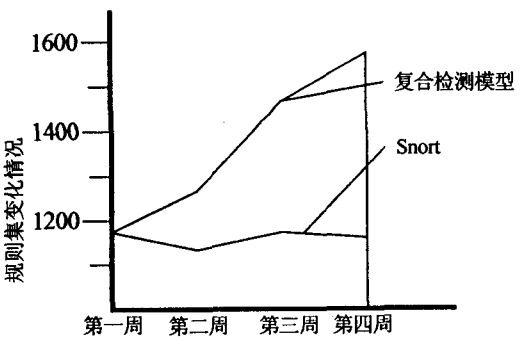

图 4 误用规则对比情况

在图 4 和图 5 中横轴是时间轴,纵轴是规则集中规则数量的变化情况,选择了四个点记录变化情况。图 4 反映 snort 中的 alert 规则的数量和复合检测模型中误用规则随时间的对比情况;图 5 反映 snort 中的 pass 规则的数量和复合检测模型中异常规则随时间的对比情况。可以发现,该框架能够有效地清洗挖掘数据,减少挖掘引擎的工作量,提取聚类数据中的频繁项形成有效规则,从而对新的攻击类型具备了一定鉴别能力。但在实验中发现系统的处理速度较慢,故在实际应用中带来了性能问题,这是下一步研究应着力解决的问题。

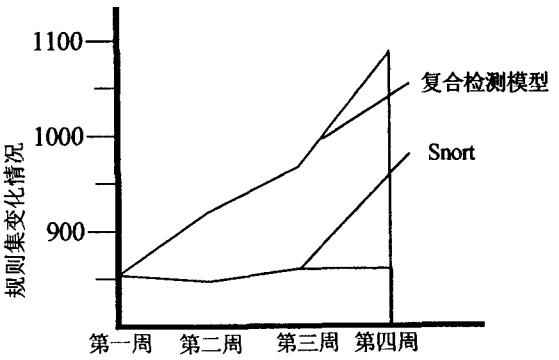

图 5 异常规则对比情况

4 结束语

在当前网络安全形势日趋复杂、网络安全状况日益严峻的情况下,作为一种主动防御技术的入侵检测技术越来越得到人们重视。针对当前入侵检测系统存在误报率和漏报率较高的问题,文中提出一种结合异常检测和误用检测两种技术、并充分利用数据挖掘优势的入侵检测模型,实验证明,该模型能自动生成新规则,对未知攻击具备一定鉴别能力。

参考文献:

[1] 卿斯汉,蒋建春,马恒太. 入侵检测技术研究综述[J]. 通信学报,2004,25(7):19-29.

[2] 王 杰,李冬梅. 数据挖掘在网络入侵检测系统中的应用[J]. 微计算机信息,2006,22(3):73-75.

[3] 朱 明. 数据挖掘[M]. 合肥:中国科学技术大学出版社,2002.

[4] Tan Pangning, Steinbach M, Kumar V. Introduction to Data Mining[M]. [s.l.]:Posts & Telecom Press,2006.

[5] 胡吉明,鲜学丰. 挖掘关联规则中 Apriori 算法的研究与改进[J]. 计算机工程与设计,2006(4):99-101.

[6] Park J S, Chen M S, Yu P S. An effective hash-based algorithm for mining association rules[C]//Proceeding of the ACM SIGMOD International Conference on Management of Data. New York: ACM, 1995:175-186.

[7] 朱桂宏,王 刚. 基于数据流的网络入侵检测研究[J]. 计算机技术与发展,2009,19(3):175-177.

[8] Bass T. Intrusion detection systems and multisensor data fusion[J]. Communications of the ACM, 2000, 43(4):99-105.