文章编号: 1673-1549(2011) 01-0077-05

基于模式识别的智能入侵检测系统

邱树伟

(汕头职业技术学院计算机系, 广东 汕头 515078)

摘 要: 文章分析了入侵检测的基本理论和技术, 结合模式识别的原理及其与入侵检测的相似性, 论证了将模式识别方法应用于入侵检测的可行性, 同时, 运用近邻法、K-近邻法和 K-均值法等关键技术设计了基于模式识别的智能入侵检测系统, 仿真实验表明, 该系统具备较高的检测率和较低的误警率。

关键词: 入侵检测系统; 人工智能; 模式识别; 近邻法

中图分类号: TP393　　文献标识码: A

随着互联网的推广, 基于网络的业务应用已深入到社会各行各业中, 网络的安全性显得格外重要。网络安全主要涵盖保密性、完整性、可用性和认证等四个方面[1]。由于网络在设计、部署和应用过程中存在缺陷, 致使网络安全确保和服务无法令人满意, 因此, 研发可信的信息安全产品已成为学术界和工业界共同努力的方向。

入侵检测技术是扩展系统安全确保能力、提高信息安全基础架构完整性的重要领域。由于入侵检测过程要面对复杂的环境和多变的攻击手段, 这就要求入侵检测系统必须具备灵活性、主动性和自适应性。基于人工智能的入侵检测技术成为人们关注的焦点[2], 特别是模式识别的应用, 更是提升入侵检测系统性能的重要方法。

1 入侵检测系统简介

入侵检测系统 (Intrusion Detection System, IDS)指入侵检测过程中所需要配置的各种软硬件的组合, 它通过对信息系统的运行状态进行实时监测, 发现各种攻击企图、攻击行为或攻击结果并作出响应, 以保证系统资源的机密性、完整性和可用性[3]。它的主要功能有: 监控、分析用户和系统的活动; 检查系统的配置和漏洞; 评估系统关键资源和数据的完整性; 识别已知的攻击行为、统计分析异常行为; 对操作系统进行日志管理; 识别违反安全策略的用户活动; 响应入侵事件等[4]。

通用入侵检测系统一般由数据采集模块、入侵分析引擎、应急处理模块及管理配置模块等组成[5]。如图 1 所示。

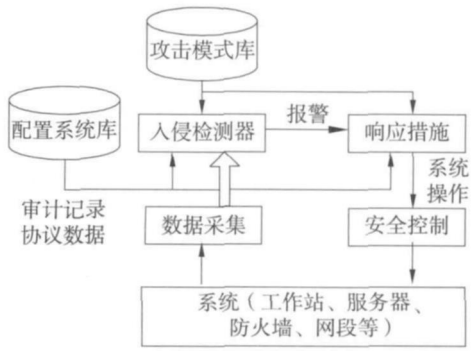

图 1 通用入侵检测系统

数据采集模块: 为入侵分析引擎提供原始数据 (如操作系统的审计日志、应用程序的运行日志和网络数据包等)。

入侵分析引擎: 对数据进行分析, 判断是否属于入侵行为并作出响应。

应急处理模块: 发生入侵后启用紧急措施 (如关闭网络服务、中断网络连接、启动备份系统等)。

管理配置模块: 为其他模块提供配置服务, 是入侵检测系统与用户的接口。

收稿日期: 2010-11-15

作者简介: 邱树伟(1979-), 男, 广东潮州人, 讲师, 硕士, 主要从事网络与分布计算、人工智能方面的研究。

2 智能入侵检测技术讨论

尽管目前许多入侵检测系统能够满足大部分用户的需求,但是,在关键领域(如金融、商务和军事等)的应用还存在若干问题[6]主要表现在:误报率偏高、报警信息过多;缺乏检测未知入侵的有效手段;自适应、自学习能力低;互操作性差,难以形成协同防御体系等。人工智能技术的运用,为解决以上问题打下了坚实的基础[7-8]。

模式识别的基本原理是:将一个输入模式与保存在系统中的多个参考模式相比较,找出最近似的参考模式,将该参考模式所代表的类名作为输入模式的类名输出。模式识别可分为学习和识别两个过程。学习是为了构造识别系统而进行的一种行为,参考模式是通过学习之后确定的。在应用识别系统的过程中,必须实时更新参考模式以增强系统的自适应性,这需要对识别结果集进行学习[9]。如图 2所示。

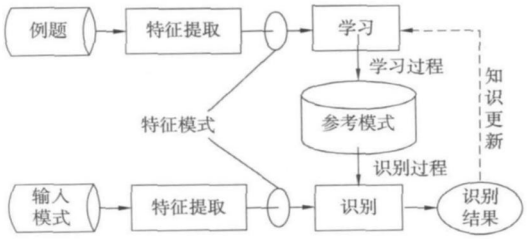

图 2 模式识别的过程

本质上,模式识别是对未知样本进行类归属判定的过程;而入侵检测也是将一个新的实例与原有的规则集进行比较归类的过程。两者工作机理非常相似。模式识别的应用对于改善入侵检测系统的识别精度、识别能力以及智能特性有着重要的影响[4]。

3 模式识别核心技术

模式识别方法有很多,本文从应用的角度出发,主要介绍近邻法[10]、k – 近邻法和 k – 均值法。近邻法是模式识别分类算法中比较常用的一种方法; k – 近邻法是基于统计的分类方法,是非参数分类的一种重要方法[11]; k – 均值法是一种基于分割的聚类方法。这三种方法是本文提出的智能入侵检测系统的核心算法。

3.1 近邻法

假设有 c 个类别 $\omega_1, \omega_2, \cdots, \omega_c$ 的模式识别问题,每类有标明的样本 N_i 个, $i = 1, 2, \cdots c$。定义 ω_i 类判别函数为:

$$g_i(x) = \min_i \| x - x_i^k \|, \quad k = 1, 2, \cdots, N_i \tag{1}$$

其中, x_i^k 的角标 i 表示 ω_i 类, k 表示 ω_i 类 N_i 个样本中的第 k 个。式 $\| x - x_i^k \|$ 表示 x 与 x_i 的欧氏距离,取欧氏距离为:

$$d = \| x - x_i \| = \sqrt{\sum_i (x - x_i)^2} \tag{2}$$

决策规则为:

$$g_j(x) = \min_i g_i(x), \quad i = 1, 2, \cdots, c \quad x \in \omega_j \tag{3}$$

对于未知样本 x, 只要比较 x 与 $N = \sum_{i=1}^{c} N_i$ 个已知类别的样本之间的欧氏距离,就可判定 x 与离它最近的样本同类。

入侵检测时,如果样本集中有一个样本 x_i 与待检样本 x 相同,即 $d = 0$, 说明此时的系统状态在样本集中有描述。这种情况类似于误用检测中的规则匹配检测法。如果样本集是准确的,那么检测结果也是准确的。如果待检样本虽然在样本集中没有明确描述,但属于某个参考样本所描述的空间范围之内,那么这个样本的属性是明确的,检测结果也是可信的。对于处于参考样本描述空间边缘或处于两个参考样本描述空间之间的待检样本,若直接用欧氏距离判定其属性,则检测结果存在一定偶然性,因此,引入 k – 近邻法来提高检测的准确性。

3.2 k – 近邻法

k – 近邻法就是取未知样本 x 的 k 个近邻,并逐一测试这 k 个近邻中的多数属于哪一类,就把 x 归为哪一类。具体为:假设这 N 个样本中,来自 ω_1 类的实例有 N_1 个,来自 ω_2 类的样本有 N_2 个, …, 来自 ω_c 类的样本有 N_c 个,若 $k_1, k_2, \cdots k_c$ 分别是 k 个近邻中属于 $\omega_1, \omega_2, \cdots, \omega_c$ 类的样本数,定义判别函数为:

$$g_i(x) = k_i, \quad i = 1, 2, \cdots, c \tag{4}$$

决策规则为:

$$g_j(x) = \max_i k_i \quad x \in \omega_j \tag{5}$$

入侵检测时,对于一个未知样本 x, 计算它与两个样本集(正常/异常行为样本集)中所有样本的欧氏距离,提取距它最近的 k 个,假如其中有 k_1 个属正常行为样本, k_2 个属异常行为样本,若 $k_1 > k_2$, 则该对象行为正常,反之,则异常。为避免 $k_1 = k_2$ 的情况出现, k 需取奇数。

3.1和 3.2的方法都存在一个问题,就是需要存储全部参考样本,当问题规模较大时,欧氏距离的计算量和存储量很大。可考虑采用两种改进的方法:

(1)对参考样本集进行组织与整理,分群分层,尽可能将计算压缩到在接近待检样本邻域的小范围内,避免全局欧氏距离计算。3.3节中的 k – 均值法将是解决方案之一。

(2)在参考样本集中挑选出对分类计算有效的样本, 使参考样本总数合理地减少, 以节约计算和存储的开销。

改进的近邻法有: 快速搜索法、剪辑法及压缩法等。

3.3 k – 均值法

在一个样本空间内, 各样本之间存在一定的距离, 距离较小的样本之间存在相似度, 即一个样本空间是 n 类具有相似度的样本集的合集。入侵检测的规则集是一个多维样本空间, 每个规则就是其中的一个样本, 同类样本之间也存在相似度, 所以, 可以考虑利用其相似性对规则集进行聚类分析。聚类分析是指根据数据的内在性质将数据分成一些聚合类, 每一聚合类中的元素尽可能具有相同的特性, 不同聚合类之间的特性差别尽可能大。

学者们提出了许多聚类算法, k – 均值法最为常用。本文采用该算法对规则集进行优化, 以达到减少计算量, 提高检测效率的目的。

k – 均值法的思想是将数据集 X 分割为 k 个聚类并使得在每个聚类中所有值与该聚类中心距离的总和最小。每个聚类的聚类中心是每个聚类的均值。算法选择的相似性度量通常是欧氏距离的倒数, 即两者的距离越小表示两者的相似度越大, 反之则相似度越小。欧氏距离公式为:

$$d_w(x_k, c_i) = \sum_{j=1}^{m} (x_{kj} - c_{ij})^2 \qquad (6)$$

该算法具体为: 把 n 个向量 $x_j (j = 1, 2, \cdots, n)$ 分 c 个类 $G_i (i = 1, 2, \cdots, c)$, 并求每个类的聚类中心, 使得非相似性 (或距离)指标的目标函数达到最小, 当选择第 i类 G_i 中向量 x_k 与相应聚类中心 c_i 间的度量为欧氏距离时, 目标函数可定义为:

$$J = \sum_{i=1}^{c} J_i = \sum_{i=1}^{c} \left(\sum_{k, x_k \in G_i} \| x_k - c_i \|^2 \right) \qquad (7)$$

这里 $J_i = \sum_{k, x_k \in G_i} \| x_k - c_i \|^2$ 是类 G_i 内目标函数。J_i 值依赖于 G_i 的几何形状和 c_i 的位置。显然, J 的值越小, 表明聚类效果越好。

在入侵检测中应用 k – 均值法, 发现取不同的初始聚类中心会产生不同的检测结果, 影响检测的稳定性。而 k – 均值法对初始聚类中心的选择是任意的, 这很可能会破坏规则集的区间相似性 (同类攻击的规则存在一定的相似性, 其区间距离也较小, 称为 “区间相似性 ”), 所以, 在运用 k – 均值法之前先采用一定的策略选择一个合适的聚类中心, 检测的效果会大大提高。

4 基于模式识别的智能入侵检测系统

根据入侵检测的基本原理, 结合模式识别的技术和思想, 设计了基于模式识别的智能入侵检测系统。其机制是: 先采用 k – 均值法对参考规则集进行聚类分析, 简化规则集的规模, 减少检测的计算量; 然后采用近邻法进行一次检测; 若无法得到理想的结果, 则改用 k – 近邻法进行二次检测。在保证检测准确率的基础上, 使系统具有一定的检测未知入侵的能力。

4.1 系统模型设计

本系统采用模块化设计, 主要包括数据采集模块, 特征提取模块, 规则处理模块, 分析检测模块和异常响应模块等。如图 3所示。

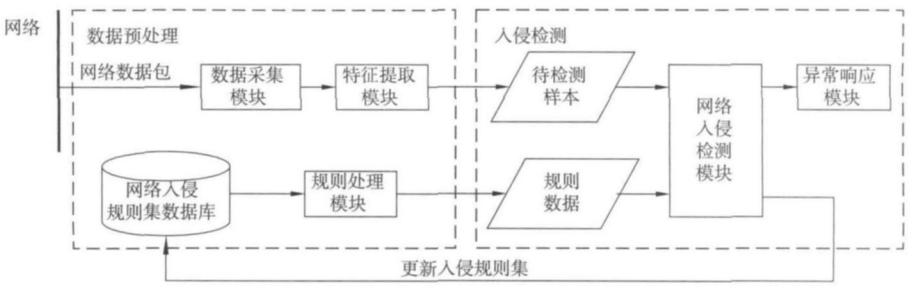

图 3　基于模式识别的入侵检测系统结构

各模块的功能如下:

数据采集模块: 实时采集网络原始数据, 并按不同的协议进行解码, 再对解码后的信息进行分片重组、流重组及代码转换等处理, 还原数据包的原始含义和数据包之间的关系。

特征提取模块: 对数据采集模块采集到的数据进行特征选择, 然后对信息进行向量化, 生成待检测样本。

规则处理模块: 进行规则集的向量化和聚类工作。首先按条读入规则, 对每条规则进行向量化处理, 得到一个规则向量集, 再对规则向量集进行聚类分析 (向量集规模较小时不必进行聚类), 生成精简的参考规则集。

分析检测模块: 是系统的核心模块。将待检测样本与参考规则集进行比较分析, 确定是否存在入侵。具体过程是: (1)采用近邻法分析待检测样本与参考规则集;

(2)若欧氏距离 d=0,即待检测样本与参考规则集中某些规则匹配,得出分析结果; (3)若 d≠0,则采用 k - 近邻法进行二次检测,得出分析结果; (4)根据分析结果判断待检测样本属正常行为或异常行为; (5)若属异常行为,立即开启异常响应措施,同时更新原规则数据库;若属正常行为,则退出。入侵检测的工作原理如图 4所示。

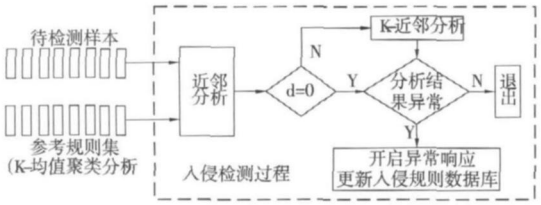

图 4 入侵检测工作原理

异常响应模块:对入侵作出响应(报警、日志记录等)。

4.2 检测算法

本系统的检测算法如下,为了突出重点,只给出关键代码。

```
Void Detect( intK)
{
for( i= 0; i< GetCount( ); i+ + ) //对于每一个待检测样本向量
{
SearchKNeighbor(K, i) //寻找该向量的 K 个近邻,记录在近邻表中
for( j= 0; j< K; j+ + ) //统计近邻表中正常和异常向量的数目
if( nt[ j]. pneighbor- > GetAttrib( ) = = 0) norm+ +; //0表示正常行为
else if( nt[ j]. pneighbor- > GetAttrib( ) = = 1) abnorm+ +; //1表示异常行为
If( norm< abnorm) alert( ); //异常则报警
}
}
```

GetCount( )表示待检测样本向量总数(正整数),SearchKNeighbor(K, i)函数用于选择第 i 个样本的 k 个近邻,GetAttrib( )函数用于提取待检测样本的特征。本系统的其他程序略。

4.3 仿真实验及性能评估

实验选用的样本数据是 KDD CUP 1999数据,来源于 MIT Lincoln Lab,它是为了评估入侵检测系统的优劣和计算入侵检测率、误测率等各项评估指标而建立的[12]。

入检测程序使用 Visual C+ + 6.0和 Matlab R2010a 编写,运行环境为: Windows Server 2003, Intel CPU 1.73GHz,内存 512M。仿真程序主要由两个模块组成: (1)预处理模块,主要用于预处理 KDD99数据集(数据清洗和属性约简等); (2)核心模块,检测算法实现及结果数据的统计和分析。仿真实验结果如图 5所示,系统平均检测率为 91.92%、误警率为 0.45%,具有较高的检测性能。

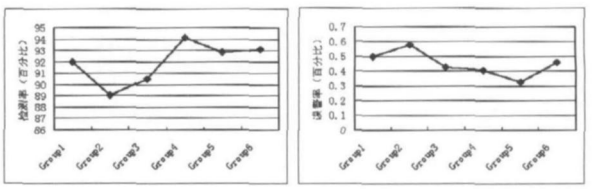

图 5 系统实验数据

系统采用模式识别中的近邻法与 k - 近邻法作为入侵检测的核心算法,其优点是检测错误率为 $p^* \leqslant p \leqslant 2p^*$, P* 为贝叶斯错误率。近邻法在 $d = \|x - x_i\| = 0$ 时达到理想检测效果,而在 $d \neq 0$时可能出现检测错误,这对参考规则集的建立提出了较高的要求; k - 近邻法由于取 k个近邻参与计算,使待检样本得到了扩充,在一定程度上降低了出错的概率。k - 近邻法中 k的选取值对检测的效率和结果影响较大,必须注意选取的方法。

近邻法和 k - 近邻法有一个共同的缺点,就是当规则集规模很大时,每次决策都需计算欧氏距离并进行比较,存储和计算开销很大(该缺点对普通入侵检测系统而言更甚),而入侵检测对准确性和实时性要求很高,需进一步克服,采用 k - 均值法对规则集进行预处理,减少了系统的时空开销,效果良好。

5 结束语

入侵检测是防火墙、数据加密和访问控制等传统安全技术的必要补充,是信息安全确保体系的重要组成部分。入侵检测系统可以对入侵行为进行识别和响应,它不仅可以检测来自网络的攻击行为,也可以监督内部用户的未授权活动。

模式识别是不断发展中的新学科,它的理论基础和应用范围正在不断发展[13]。本文提出了把模式识别方法应用到入侵检测中,将入侵检测问题转化为模式识别问题来处理,是一种较有价值的解决方案。基于模式识别的入侵检测系统自适应/学习能力强、成本低和健壮性好,能有效提高系统的安全性。

但是,本系统仍存在缺陷:为保证参考规则集的有

效性和实时性,需要提取海量的对象行为特征;在高带宽的网络环境下,为缩短检测响应时间,对检测算法的时空效率提出更高的要求[14]。这两点对入侵检测系统的效能来说具有决定性意义,如何快速构建入侵参考模式知识库、进一步提高检测算法的智能性和效率,将是进一步研究的方向。

参考文献:

[1] 郭敬宇. 一种用于入侵检测的改进的动态机器学习模型[J]. 计算机工程, 2004, 30(14): 103-104, 145.

[2] 闫巧,毛晓波,闫戈林. 人工智能在入侵检测系统中的应用[J]. 计算机工程与应用, 2002, 20: 56-59.

[3] 杨孔雨,王秀峰. 计算智能及免疫理论在入侵检测中的应用研究[J]. 计算机应用研究, 2004(5): 152-154.

[4] 黄剑峰. 基于模式识别算法的入侵检测研究[D]. 南京: 南京理工大学, 2007.

[5] Hart P E. The condensed nearest neighbor rule[J]. IEEE Transactions on Information Theory, 1968, 14: 515-516.

[6] 郑凯元. 基于计算智能的自主网络入侵检测方法研究[D]. 成都: 电子科技大学, 2009.

[7] 魏大庆,欧阳俊林. 基于数据挖掘的网络入侵检测模型研究[J]. 四川理工学院学报: 自然科学版, 2006, 19(5): 59-61.

[8] Brighton H, Mellish C. Advances in instance selection for instance-based learning algorithms[J]. Data Mining and Knowledge Discovery, 2002, 6: 153-172.

[9] 沟口理一郎,石田亨. 人工智能[M]. 北京: 科学出版社, 2005.

[10] 胡煜. 主分量分析法和K-近邻法应用于基因芯片数据分析[J]. 北华大学学报: 自然科学版, 2008, 9(1): 12-15.

[11] 颜辉. K-近邻法在入侵检测中的应用[J]. 吉林工程技术师范学院学报: 工程技术版, 2003, 19(12): 19-22.

[12] 简清明,曾黄麟,叶晓彤. 粗糙集特征选择和支持向量机在入侵检测系统中的应用[J]. 四川理工学院学报: 自然科学版, 2009, 22(5): 62-64.

[13] 蔡自兴,徐光祐. 人工智能及其应用[M]. 北京: 清华大学出版社, 2004.

[14] 赵丽萍. 基于模式识别的入侵检测模型[J]. 电脑开发与应用, 2008, 21(6): 46-48.

Intelligent Intrusion Detection System Based on Pattern Recognition

QIU Shu-wei

(Computer Department, Shantou Polytechnic, Shantou 515078, China)

Abstract The basic theory and technology of intrusion detection were introduced, combined the principles of Pattern Recognition with the similarity of Pattern Recognition and intrusion detection, the Pattern Recognition applied to the feasibility of intrusion detection system was demonstrated, at the same time, the intelligent intrusion detection system based on Pattern Recognition was designed using nearest neighbor, k-nearest neighbor and k-means, simulation results show that it has a higher detection rate and lower false alarm rate.

Key words intrusion detection system; Artificial Intelligence; Pattern Recognition; nearest neighbor