

基于 Agent 和聚类分析的网络入侵检测研究

赵振辉

(天津市信息中心, 天津 300201)

摘 要: 为了解决现有的入侵检测系统响应网络环境慢、需要较多人工干预的局限,采用智能体技术和聚类分析设计一种新的网络入侵检测方法.使用智能体技术实现入侵检测系统的分布式设计,利用分层的控制智能体实现入侵检测系统的自主控制,同时使用基于模糊 C 均值算法的数据挖掘技术对网络数据进行检测分析,并利用加权算法对模糊 C 均值算法进行改进,提高系统的检测能力.结果表明,该系统能够减少人工干预,对网络环境响应较快,入侵检测性能也得到了提高.

关键词: 入侵检测系统;智能体;数据挖掘;模糊 C 均值算法;加权算法

中图分类号: TP393

文献标识码: A

文章编号: 1000-7180(2013)03-0165-04

Research on Network Intrusion Detection Based on Agent and Clustering Technology

ZHAO Zhen-hui

(Tianjin City Information Center, Tianjin 300201, China)

Abstract: In order to solve the limitations of slow network response and more manual intervention in the existing intrusion detection system, a new method of network intrusion detection is proposed based on agent technology and clustering analysis. The distributed intrusion detection system is designed with the agent technology, in which the system is self-controlled by the hierarchical control agent, and the data analysis of network is achieved through data mining technique based on fuzzy C-means algorithm (FCM). The weighted algorithm is used in the FCM algorithm for the improvement of the system detection capabilities. The results show that the system is able to reduce manual intervention, achieve faster response for the network environment, and improve the performance of the intrusion detection.

Key words: intrusion detection system; agent; data mining; fuzzy C-means algorithm; weighting algorithm

1 引言

目前计算机网络的应用已遍布生活工作的各个方面,随着 Internet 网络的迅速发展,计算机网络安全成为影响国家、企业生存发展的重要问题,受到了人们的普遍重视^[1].如今防火墙、加密解密等网络安全技术已得到广泛应用,入侵检测是一种新颖有效的网络安全技术,能够积极主动地对网络进行安全防护,因此得到了广泛应用和深入研究.但现有的入侵检测系统普遍需要大量人工干预,响应网络变化

的反应较慢,检测性能也需要提高^[2].本文提出了一种基于智能 Agent 和加权 FCM 算法的网络入侵检测技术提高检测系统的自主性和检测性能.

2 入侵检测

入侵检测是指为检测计算机网络入侵行为所采取的对策,在计算机网络或者系统的某些关键结点提取网络信息并进行统计分析,检测所得信息中是否存在违背安全规则的异常现象.入侵检测系统是实施入侵检测的软件和硬件工具,自动监控网络状

收稿日期: 2012-06-15; 修回日期: 2012-07-18

基金项目: 国家自然科学基金项目(60874582)

态并进行入侵分析^[3].

通用的入侵检测系统模型如图 1 所示. 主体活动是指导包括系统资源在内的与主体相关的活动对象在被检测系统上的活动, 审计记录是探测器提供给分析引擎的分析数据, 活动简档保存主体正常运行的相关信息, 异常记录是入侵等异常事件发生的情况^[4].

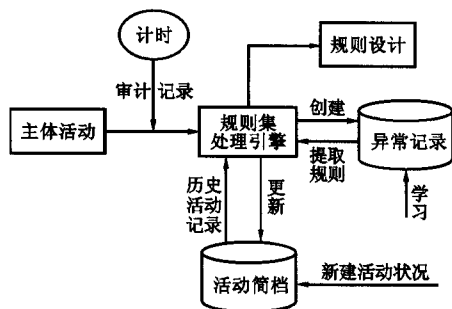


图 1 通用的入侵检测系统模型

入侵检测过程的基本步骤包括信息收集、信息分析、信息响应. 信息收集基于主机的信息数据、基于网络的信息数据和无线资源派生数据, 收集的主要对象是活动日志, 经处理后形成审计数据. 信息分析是入侵检测的核心步骤, 分为构建分析器、分析实际现场数据和反馈提炼三个过程. 信息响应是对检测发现的入侵行为进行反应, 如进行告警的被动响应和对入侵者采取反击行为、提升系统抵抗能力的主动响应. 在如今网络数据急剧膨胀、网络安全威胁日益增大的情况下, 入侵检测系统需要有快速响应环境变化而无需人工干预的智能灵活的反应能力, 智能 Agent 是一种较好的解决方式.

3 基于 Agent 的网络入侵检测

Agent 指能在一定环境中能够持续自主运行的软件实体, 能够根据环境的变化不需要人工控制自动做出对应的反应. 在理想状态下, Agent 应当具有从过去的活动中学习的能力, 同一环境下的 Agent 也应当能相互合作^[5]. Agent 有显著的特征: (1) 反应性, 有选择地反应的能力; (2) 自主性, 目标驱动的自动应变能力; (3) 全作性, 多 Agent 合作的能力.

Agent 引入入侵检测系统之后, 系统可以根据网络环境的变化自动适应并不断地学习, 并且可以实现多 Agent 的分布式网络, 每个 Agent 可以独立完成某项功能, 并可以动态地删除和添加. 这种分布式入侵检测系统将系统进行分层, 下层部件受上层控制, 并由管理中心统一管控, 效率较高, 其余系统模型如图 2 所示.

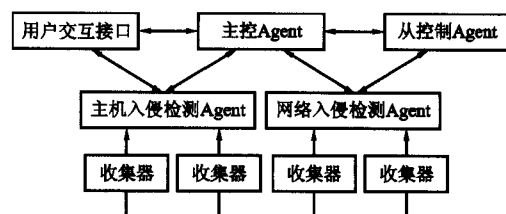


图 2 多 Agent 检测系统结构图

其中用户交互接口是用户与系统交互的图形界面管理接口, 系统可将系统运行状态信息显示给用户, 用户可通过这个接口设置配置信息. 主控 Agent 是检测系统的核心部位, 向系统内所有的主机和 Agent 发送控制命令, 并分析其下的 Agent 的报警信息进行响应. 从控制 Agent 是主控 Agent 的备用配置, 当主 Agent 受到攻击瘫痪时, 从控制 Agent 将代替主控 Agent 运行. 主机入侵检测 Agent 配置在各服务器上, 监控各服务器, 当发现入侵后上报控制 Agent 由控制 Agent 做出响应. 网络入侵检测 Agent 设布在网关或者路由上, 监控网段信息, 若发现可疑数据流, 上报控制 Agent, 由控制 Agent 做主要的综合分析判定.

多 Agent 入侵检测系统的网络呈树形结构, 重要部位使用 Agent, 可在较少人工控制的情况下自主监控网络并能处理突发事件. 但是在进行入侵检测的过程中, 需要对搜集器上报的数据进行分析处理, 这种分布式网络就需要处理大量的网络信息数据, 对分析处理引擎提出了严峻的挑战, 选择合适的分析处理算法对多 Agent 入侵检测系统的效率和性能有十分重要的意义, 基于加权模糊 C 均值算法的数据挖掘技术是一种较优的选择.

4 改进的 FCM 算法在入侵检测中的应用

4.1 FCM 算法的基本原理

FCM 算法是利用隶属度确定各数据点的聚类归属的聚类算法^[6], 主要是将特征空间 $X = \{x_1, x_2, \dots, x_n\}$ 分为 C 个模糊组, 用模糊隶属矩阵 $U = [u_{ij}] \in R^n$ 表示, 其中 u_{ij} 表示第 j ($j = 1, 2, \dots, n$) 点对于类 i ($i = 1, 2, \dots, c$) 的隶属度, 对 u_{ij} 的要求为:

$$\begin{cases} \sum_{i=1}^c u_{ij} = 1, \forall j = 1, 2, \dots, n \\ \sum_{j=1}^n u_{ij} > 0, \forall i = 1, 2, \dots, c \\ u_{ij} \in [0, 1], \forall i, j \end{cases} \quad (1)$$

通过将算法的目标函数推广可得到更通用的情况

$$J_m(U, c_1, \dots, c_i) = \sum_{i=1}^c \sum_{j=1}^n (u_{ij})^2 (d_{ij})^2 \quad (2)$$

式中, $J_m(U, c_1, \dots, c_i)$ 是目标函数, $c_i \in R^n$ 是第 i 个模糊组的聚类中心, $d_{ij} = \|x_j - c_i\|$ 是第 i 个聚类中心和第 j 个点的欧氏距离, 且有

$$d_{ij}^2(x_j, c_i) = (x_j - c_i)^T A (x_j - c_i)$$

$J_m(U, c_1, \dots, c_i)$ 是目标函数, 由各点到对应的中心的距离平均和, $m \in [1, \infty]$ 是模糊加权因子, 用以表征隶属矩阵的模糊程度。

当 u_{ij} 满足式(1) 时, 可得式(2) 最优化的必要条件:

$$c_i = \frac{\sum_{j=1}^n (u_{ij})^m x_j}{\sum_{j=1}^n (u_{ij})^m} \quad i = 1, 2, \dots, c \quad (3)$$

$$u = \frac{1}{\sum_{k=1}^c \left(\frac{d_{ij}}{d_{kj}} \right)^{\frac{2}{m-1}}} \quad I_j = \phi \quad (4)$$

式中 $I_j = \{i \mid 1 \leq i \leq c, d_{ij} = 0\}$. 参量 m 是控制算法的主要参数, 用来控制分类矩阵 U 的模糊程度, m 越大越模糊. 这样 FCM 形成了迭代过程, 经过多次迭代运算后, 取得 $J_m(U, c_1, \dots, c_i)$ 的最小值。

4.2 FCM 算法实施入侵检测的方法

FCM 算法主要有两个参量, 即模糊指数 m 和聚类数目 c , c 是大于 1 的整数, 但远小于聚类总样本数, m 是控制样本的主要参数^[7], 取值范围是 $[1.5, 2.5]$, 在入侵检测中取中间值 2. FCM 算法在入侵检测中的实现过程为:

(1) 设置 c 和 m , 初始化聚类中心向量 $c_i, 2 \leq i \leq c$, 设置误差阈值 ϵ ;

(2) 计算数据与聚类中心的距离 d_{ij} , 按式(4) 算出隶属度的值;

(3) 按式(2) 计算代价函数 J_m , 并按式(3) 修正聚类中心 c_i ;

(4) 计算代价函数的变化量 ΔJ_m , 如果 $\Delta J_m \geq \epsilon$ 表明不收敛, 转向(2) 继续计算, 否则算法结束。

算法输出结果是目标函数值、聚类中心向量矩阵和隶属度模糊矩阵, 入侵检测时, 隶属度模糊矩阵是分析对象, 根据最大隶属原则, 得到各聚类的数据值按大小排序, 根据预先设定的中心值 m , 大于 m 的聚类簇判定为正常类, 否则判定为异常类。

4.3 FCM 算法实施入侵检测的流程及可行性分析

在实际工作中, 基于 FCM 算法的入侵检测主要分为三个工作阶段: 数据预处理、检测分析阶段和决策阶段. 其算法流程图如图 3 所示。

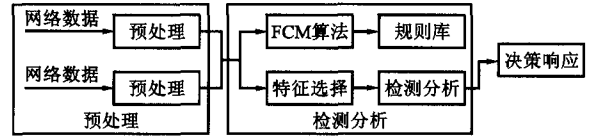


图3 基于 FCM 算法的入侵检测流程图

其中的数据预处理的工作是将网络中数据交互记录和各计算机内部的日志信息进行压缩转换等预处理工作, 以统一格式存储, 其中网络数据包是分析的重点, 将被还原成基于传输层的连接记录, 提取网络间计算机通信信息, 作为可以判定的特征属性. 这样, FCM 算法可将处理好的数据进行入侵检测, 并且异常入侵检测率较高. 对以往数据进行统计分析可以得到合适的阈值 m , 使入侵检测的误检测率较低, 而且 FCM 算法本身具有较好的自适应特性和扩充性, 可增加检测的即时性, 减少修改增加人工代码的麻烦. 但是由于 FCM 算法本身的局陷性, 异常检测的能力还有不足, 因此利用加权算法改进关联规则等在异常检测的不足。

4.4 改进的 FCM 算法实施入侵检测的方法

在 FCM 算法中, 数据属于某一聚类是根据数据与该聚类中心的欧氏距离来决定的, 并利用隶属度进行属于某个类的程度量化. 但隶属度是相对数量, 对一些典型性的异常数据的判定有一定的偏差, 因此导致了 FCM 算法在异常检测中有一定的不足。

针对这一缺陷, 设定对被判定的数据信息对某个聚类的加权值控制算法, 使隶属度大的数据对聚类中心的影响变大, 隶属度小的数据值对聚类中心位置的影响变小, 将其应用到数据的更新中, 则式(3) 变为

$$c_{ij} = \frac{\sum_{j=1}^n (u_{ij})^m x_{ij}}{\sum_{j=1}^n (u_{ij})^m} \quad i = 1, 2, \dots, c \quad (5)$$

对 u_{ij} 赋予一个权值 c_{ij} , 在式(5) 的基础上减去 $(1 - u_{ij})/2$, 这样使得 FCM 算法的隶属度是 1 时, 加权后仍为 1, 隶属度为 0 时加权后仍为 0, 并且隶属度小的数据对象对聚类中心的影响度变小, 权值为:

$$a_{ij} = u_{ij} - \frac{1 - u_{ij}}{2} \cdot u_{ij} \quad i = 1, 2, \dots, c; \quad j = 1, 2, \dots, n \quad (6)$$

新的聚类中心变为

$$c_{ij}' = \frac{\sum_{j=1}^n (a_{ij})^m x_{ij}}{\sum_{j=1}^n (a_{ij})^m} \quad i = 1, 2, \dots, c \quad (7)$$

聚类中心变化的实质是 FCM 算法在入侵检测的基础上,在过程(3)中设置新的聚类中心,其余步骤不变,这样可以消除孤立点对聚类中心的影响,使其更接近最佳点.

4.5 实验仿真和结果分析

为验证改进的 FCM 算法在入侵检测中的应用效果,使用 Matlab 进行仿真检验,数据源使用 KDD Cup 1999 网络数据集,这是麻省理工学院创建的为入侵检测效果评估而设立的测试数据集,数据来源于美国空军局域网的仿真环境,包含了 7 个周的网络数据流量,有约 5 百万条网络连接记录^[8]. 仿真试验所用的入侵数据格式主要由个部分组成:连接时间、连接 TCP/UDP 属性类别、http/ftp/smtp 服务类型类别、连接标记、记录参数和连接安全性标记,连接安全性标记说明该连接是正常连接还是非法入侵连接. 仿真所选取的 4 个数据集结构如表 1 所示.

表 1 仿真数据集结构

	实例数	正常实例数	入侵实例数	入侵实例数	分类
数据集 1	3100	3000	100	13	4
数据集 2	3100	3000	100	13	3
数据集 3	3100	3000	100	7	3
数据集 4	3100	3000	100	18	4

在 Matlab 中利用 FCM 算法和改进的 FCM 算法对这四组数据集进行入侵检测的仿真,仿真结果用入侵检测率 P_i 和误检率 P_m 进行定量分析,入侵检测率 P_i 表示检测出的入侵行为数目占总入侵数目的比例,误检率 P_m 表示将非入侵行为判定为入侵行为的比例,计算方法为:

$$P_i = \frac{N_i}{N_e} \tag{8}$$

$$P_m = \frac{(N_m - N_i)}{N} \tag{9}$$

式中, N_i 、 N_m 、 N_e 、 N 分别表示检测入侵行为数、判定入侵行为数、入侵实例数、实例总数. 得出的数值结果如表 2 所示.

表 2 仿真结果比较

	入侵检测率		误检率	
	FCM 算法	改进的 FCM 算法	FCM 算法	改进的 FCM 算法
数据集 1	60%	69%	3.293%	1.378%
数据集 2	67%	73%	3.411%	1.605%
数据集 3	54%	66%	2.681%	1.576%
数据集 4	71%	76%	3.531%	1.832%
平均值	63%	71%	3.229%	1.672%

从表 2 可以看出,在入侵检测率上,FCM 算法的平均检测率为 63%,而改进后的 FCM 算法的检测率为 71%,入侵检测率有了较大的提高,与此同时,改进后的算法的平均误检率从 3.229%降到了 1.672%,性能得到了较大的提升.

5 结束语

基于多 Agent 分布式结构的入侵检测系统可在较少人工控制的情况下自主监控网络并能处理突发事件,系统扩展性、灵活性得到了较大提升,但也提高了网络信息数据分析的难度. 基于加权 FCM 算法的数据挖掘技术可以及时快速处理大量的网络信息数据,并且检测率和误检率也有了一定程度的提高,可靠性也得到了保证. 但是对日益严峻的网络安全形势仍有待进一步提高检测率,这是下一步研究工作的重点.

参考文献:

[1] 孙强. 基于定量安全风险评估模型的网络安全管理平台[J]. 微电子学与计算机, 2010, 27(5): 70-73.

[2] Garcia P Teodoro. Anomaly-based network intrusion detection: techniques, systems and challenge [J]. Computers & Security, 2009, 28(1): 18-28.

[3] 彭道刚,杨平,韩璞,等. 基于并行自学习的神经网络内模控制及其应用研究[J]. 计算机仿真, 2004, 5(3): 203-206

[4] 刘文涛,胡家宝. 基于排挤遗传算法的入侵检测方法[J]. 计算机工程与应用, 2010, 46(33): 91-95.

[5] L Xiao, X Liao, H Wang. Cluster consensus on discrete-time multi-agent networks[J]. Computer Science and Service System, 2011: 2538-2542.

[6] 王丹,吴孟达. 粗糙模拟 C 均值整合聚类[J]. 国防科技大学学报, 2011, 33(3): 337-342.

[7] 关庆,邓赵红,王士同. 改进的模糊 C 均值聚类算法[J]. 计算机工程与应用, 2011, 47(10): 27-29.

[8] Mohammad Sazzadul Hoque. An implementation of intrusion detection system using genetic algorithm[J]. International Journal of Network Security & Its Applications, 2012, 4(3): 109-120.

作者简介:

赵振辉 男, (1975—), 工程师. 研究方向为计算机网络.