

基于聚类粒子群算法网络异常检测模型研究

牛红惠, 徐 甜

(安阳师范学院 计算机教学部, 河南 安阳 455000)

摘 要: 提出了一种新的基于聚类算法和遗传算法相结合的入侵检测方法模型. 算法对聚类的中心采用二进制编码, 将网络的正常行为和非正常行为分为不同的类, 把每个点到它们之间的各自的聚类中心的欧几里得距离的综合作为相似度量, 然后采用粒子群优化算法, 有效的降低网络拓扑路径长度, 通过优化算法来寻找聚类的中心. Matlab 仿真实验结果表明, 提出的改进的网络异常检测方法, 与较传统网络入侵检测系统模型相比, 具有更好的入侵识别率和检测率, 同时提高了算法的执行效率.

关键词: 网络异常; 均值聚类; 遗传算法; 入侵检测

中图分类号: TP319

文献标识码: A

文章编号: 1000-7180(2012)03-0102-04

Model of Intrusion Detection Based on Means Clustering and PSO Algorithm

NIU Hong-hui, XU Tian

(Anyang Normal University of Computer Center, Anyang 455000, China)

Abstract: A new clustering algorithm and genetic algorithm based on a combination of intrusion detection model. Algorithm to cluster the center of the binary code of each point to their respective cluster centers between the Euclidean distance as similarity measure integrated, and then use genetic algorithm, effectively reduce the path length of the network topology, the genetic algorithm to find the cluster center. Matlab simulation results show that the improved network anomaly detection, and the more traditional model of network intrusion detection system, compared with the invasion of a better recognition rate and detection rate, while increasing the efficiency of the algorithm.

Key words: network anomaly; means clustering; genetic algorithm; intrusion detection

1 引言

网络技术高速发展的今天, 人们越来越依赖于网络进行信息的处理. 因此网络安全就显得相当重要, 随之产生的各种网络安全技术也得到了不断地发展. 入侵检测是对入侵行为的检测.

随着机器学习算法已逐渐在入侵检测系统中应用广泛, 特别是基于神经网络的入侵检测算法, 由于其独特的优点, 被广泛应用在入侵检测系统中. 随着网络攻击技术的不断提高和网络安全漏洞的不断发现, 传统防火墙技术加传统 IDS 的技术, 已经无法

应对某些安全威胁^[1].

针对入侵检测系统存在的比较高的漏报率以及高的误报率, 同时也存在入侵检测的数据存在维数大、冗余度高等缺陷, 针对入侵检测系统存在的比较高的漏报率和误报率, 基于均值聚类和遗传算法的优点, 提出了一种新的基于聚类算法和遗传算法相结合的入侵检测方法模型. 算法对聚类的中心采用二进制编码, 把每个点到它们之间的各自的聚类中心的欧几里得距离的综合来作为相似度量, 然后采用遗传优化算法, 有效的降低网络拓扑路径长度, 通过遗传算法来寻找聚类的中心. 通过仿真实验对其

收稿日期: 2011-07-15; 修回日期: 2011-08-20

基金项目: 河南省科技厅科技计划项目(112300410129)

进行验证.具有更好的入侵识别率和检测率,同时提高了算法的执行效率.

2 网络异常检测原理

所谓的网络入侵检测主要目的就是针对正在入侵进行不断识别的过程.对网络系统中的关键点进行信息的收集并分析,从而发现是否有一些信息违反了网络安全策略的一些行为以及攻击的迹象.如图1所示的是网络异常检测工作关系原理图.

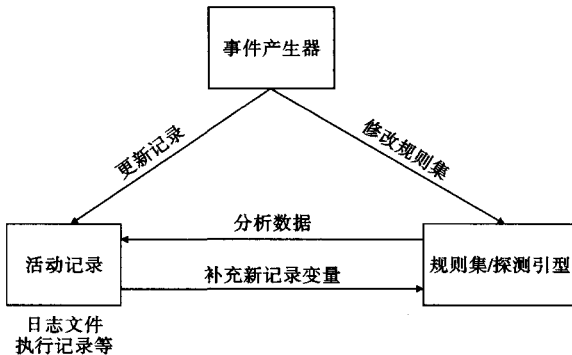


图1 网络异常检测工作关系图

3 聚类粒子群算法网络异常检测

3.1 粒子群算法

传统的粒子群优化算法(PSO)^[2]就是一种基于进化计算技术方法,它主要是由两位美国博士发明得来的.主要源于对群鸟的捕食行为的研究.粒子群算法和传统的遗传算法比较相似,也是一种基于迭代的优化工具算法.该方法首先初始化为一组随即的解,然后通过迭代搜寻最优值.然而其并不是遗传算法采用的交叉变异操作.和传统的遗传算法相比,粒子群算法的优点在于算法比较简单,而且没有过多的参数需要调整.

粒子群算法数学表达式可以采用下面的公式来表示粒子每轮的更新行为:

$$V_{id}(t+1) = V_{id}(t) + C_1 \times \Phi_1 \times (P_{bid} - X_{id}(t)) + C_2 \times \Phi_2 \times (P_{gid} - X_{id}(t)) \quad (1)$$

$$X_{id}(t+1) = X_{id}(t) + V_{id}(t+1) \quad (2)$$

其中, V_{id} 表示的是粒子的速度, X_{id} 表示的为粒子的位置, C_1 和 C_2 表示的是学习因子或者是加速常量, ϕ_1, ϕ_2 都在 $(0, 1)$ 区间内两个随机正数, p_{bid} 为个体意识(个体最佳解位置), p_g 表示的是群体最佳解位置.

(1) 初始化,设定加速常数 C_1 和 C_2 ,最大进化代数 T_{max} 将当前进化代数置为 $t=1$,在定义空间 R^n

中随机产生 m 个粒子 $X_1, X_2, \dots, X_n$,组成初始种群 $X(t)$,随机产生各粒子初速度 $V_1, V_2, \dots, V_s$ 组成位移变化矩阵 $V(t)$.

(2) 计算种群中所有粒子的适应值,初始化每粒子 $pbest$ 为当前粒子,设定种群的 $gbest$ 为当前种群的最优粒子.

(3) 种群 $X(t)$ 演化,对种群中的每一个粒子:

① 按上式更新粒子的位置和速度.

② 计算粒子的适应值.

③ 比较粒子的适应值和自身的最优值 $pbest$.如果当前值比 $pbest$ 更优,则更新 $pbest$ 为当前值,并更新 $pbest$ 位置到 n 维空间中的当前位置.

④ 比较粒子适应值与种群最优值.如果当前值比 $gbest$ 更优,则更新 $gbest$ 为当前种群最优粒子.

(4) 检查结束条件,如果满足,那么结束寻优;否则,置 $t=t+1$,转至第二步骤.结束条件为寻优达到最大进化代数 T_{max} ,或评价值小于给定精度 ϵ .

3.2 C-均值聚类算法

模糊 C-均值聚类(Fuzzy c-means clustering)^[8]结果是用隶属度(membership)表示,表明属于某种类型的程度.它是目前唯一的重叠分类方法,模糊 C-均值聚类是以所分组内平方和最小化为判据,组内平方和定义为^[3]:

$$J.(U, V, A) = \sum_{i=1}^N \sum_{j=1}^C (U_{ij})^m (d_{ij}^2 A)^2 \quad (3)$$

$$d_{j,k}^F(p) = \begin{cases} d_{j,k}^F(p), & \text{若 } C_{j,k}^A(p) > C_{j,k}^B, p \in r_i^F \\ d_{j,k}^B(p), & \text{若 } C_{j,k}^A(p) \leq C_{j,k}^B, p \in r_i^F \end{cases} \quad (4)$$

其中 N 表示的是样方数; C 表示的是预分组数.

$(d_{ij}^2 A)^2 = \|X_i - V_j\|^2 A = (X_i - V_j)^T A (X_i - V_j)$ $(d_{ij}^2 A)^2$ 表示的是距离指标^[4], $U = \{U_{ij}\}$ 表示的是隶属度矩阵, U_{ij} 指的是第 i 个样方在第 j 个分组中的隶属度, V_j 表示的是第 j 类的中心, X_i 表示的是第 i 个样方的基础值,一般为—排序坐标值; m 表示的是模糊参数 $(1 \leq m < \infty)$; C 为要分的组数,是预先确定的; A 为任一 N 维矩阵,是人为给定的,以使数据保持一定的空间结构.如果 A 取单位矩阵,则式(2)就简化为^[5-6]:

$$d_{ij}^2 = \|X_i - V_j\|^2 \quad (5)$$

模糊参数 m 可以取大于或等于 1 的任何值.

3.3 粒子群优化均值聚类

当输入样本 X_i 和聚类的数目 c ,将每个样本随机的指派为某一类别并作为最初的聚类划分,计算

聚类的中心位置,作为初始粒子的位置编码,然后算法计算粒子的适应度,通过反复迭代 N 次,一共生成了 N 个初始化粒子群^[7]

For $t=1$ to $\max_Generations$

接下来依据粒子的聚类中心的编码,按照最近邻规则,来确定对粒子进行聚类划分.依照聚类的划分来重新计算新的聚类的中心位置,并更新粒子的适应度值,最后算法输出最优的位置编码,确定聚类的中心位置.

在采用粒子群算法过程中,对新一代个体进行划分类别时,可能会出现空的聚类,若出现了空聚类,那么可以从其他的非空聚类中提取距离聚类中心最远的向量,把这个向量放进空的聚类中.

4 实验结果与分析

本文所有的实验都是在 PC P4 T2310 1.86G, 2GRAM, Intel82865G 显卡的计算机上做了实验,实验环境为 MATLAB7.0. 本文所有的仿真实验数据都来自于网络入侵检测领域比较权威的 KDD CUP 99 数据集.该 KDD CUP 99 数据集中主要包含了许多正常数据与多种异常数据,总共包括六大类的网络入侵攻击行为:其一是拒绝服务攻击(DoS);其二是非授权得到超级用户权限攻击(U2R);其三是远程非授权访问攻击(R2L);还有就是漏洞探测与扫描攻击(PROBE).该数据局 KDD CUP 99 共包含 5007894 条原始数据,采用 SQL SERVER 2000 数据库技术对重复记录进行删

除,其获得 1050454 条有效数据,其分布为:正常有 812,814 条,DoS 攻击有 247267 条,R2L 攻击有 999 条,U2R 攻击有 52 条,PROBE 攻击有 13860 条.

本文首先采用只包括一种正常数据集作为训练数据,然后针对这些异常入侵数据集进行检测,并计算出 X_i 和各个聚类的中心距离 d ,当距离小于事先设定的阈值时,那么此时 X_i 表示的是正常的数据,否则认为是异常数据.本文选择不同的聚类数目 C ,通过设定不同的聚类数,来测试比较四种入侵数据类型的不同的检测率,并比较了传统的聚类入侵检测算法和标准遗传算法入侵检测模型,实验结果如表 1 所示.在原始数据中包括了两种属性,一个是离散的,另一个是连续的特征,试验中不需要对离散的特征进行预处理,对于连续属性特征,不同的属性特征具有各不相同的度量标准,此时的数据属性特征会被遮掩,为了处理好该问题,所以必须要将数据的特征属性值进行标准化:

(1) 对平均的绝对偏差 S 进行计算

$$S = \frac{1}{n} \sum_{i=1}^n (X_i - m) \quad (6)$$

其中, X_i 表示的是数据样本, m 表示的是样本均值.

$$m = \frac{1}{n} \sum_{i=1}^n X_i \quad (7)$$

(2) 标准化的数据计算

$$Y_i = \frac{X_i - m}{S} \quad (8)$$

表 1 不同数据攻击三种算法比较结果

攻击类型	均值聚类			遗传算法			本文算法		
	检测率	正确率	漏报率	检测率	正确率	漏报率	检测率	正确率	漏报率
DoS	87.2	59.2	12.1	78.1	55.0	21.5	91.2	63.6	7.4
PROBE	88.6	68.1	10.5	79.3	60.2	22.4	95.2	70.8	6.0
U2R	21.2	16.1	80.5	19.0	14.6	79.8	19.9	16.6	79.1
R2L	24.9	19.3	73.8	23.2	20.1	76.9	26.9	23.1	73.0
混合攻击	14.8	11.9	83.1	14.0	10.6	85.8	16.4	13.1	84.2
新攻击	10.6	10.3	88.0	9.0	7.9	90.0	12.7	9.6	88.1

从表 1 的测试结果可以看出,本文提出的均值聚类粒子群优化算法网络检测时间少,收敛速度快,对明显提高了各种网络攻击的检测率和正确率,漏报率较低.证明了提出的网络入侵检测性能得到明

显的改善,提高网络入侵检测系统总体的检测性能.

5 结束语

本文研究了网络安全入侵检测问题.针对入侵

检测系统存在的比较高的漏报率以及高的误报率,同时也存在入侵检测的数据存在维数大、冗余度高等缺陷,本文基于均值聚类和遗传算法的优点的基础上,提出了一种新的基于聚类算法和遗传算法相结合的入侵检测方法模型.算法对聚类的中心采用二进制编码,将网络的正常行为和非正常行为分为不同的类,把每个点到它们之间的各自的聚类中心的欧几里得距离的综合来作为相似度量,然后采用粒子群优化算法,有效的降低网络拓扑路径长度,通过优化算法来寻找聚类的中心.仿真实验结果证明了本文提出的改进的网络异常检测方法,与较传统网络入侵检测系统模型相比,具有更好的入侵识别率和检测率,同时提高了算法的执行效率.

参考文献:

- [1] 段智彬,孙恩昌,张延华,等. 基于 ARMA 模型的网络流量预测[J]. 中国电子科学研究院学报, 2009,4(4): 352-356.
- [2] 王俊松,高志伟. 基于 RBF 神经网络的网络流量建模及预测[J]. 计算机工程与应用, 2008,44(13):7-9.
- [3] 冯海亮,陈涤,林青家,等. 一种基于神经网络的网络流量组合预测模型[J]. 计算机应用, 2006,26(9): 2206-2208.
- [4] Lahiri S K, Ghanta K C. Prediction of pressure drop

of slurry flow in pipeline by hybrid support vector regression and genetic algorithm model [J]. Chinese Journal of Chemical Engineering, 2008,16(6): 841-848.

- [5] Eslam Pourbasheer, Siavash Riahi, Mohammad Reza Ganjali, et al. Application of genetic algorithm-support vector machine (GA-SVM) for prediction of BK-channels activity[J]. European Journal of Medicinal Chemistry, 2009, 44(12):5023-5028.
- [6] Huang Cheng-Lung, Wang Chieh-Jen. A GA-based feature selection and parameters optimization for support vector machines [J]. Expert Systems with Applications, 2006,31(2):231-240.
- [7] Lin Shih-Wei, Ying Kuo-Ching, Chen Shih-Chieh, et al. Particle swarm optimization for parameter determination and feature selection of support vector machines [J]. Expert Systems with Applications, 2008,35(4): 1817-1824.

作者简介:

牛红惠 女,(1972—),硕士,讲师.研究方向为神经网络算法与应用、计算机网络.

徐甜 女,(1968—),硕士,副教授.研究方向为数据挖掘算法与应用、人工智能.

(上接第 101 页)

- [2] 谭晓玲,许勇,张凌,等. 网络流量短期预测方法的研究与应用[J]. 计算机工程与设计, 2006, 27(8): 1341-1345.
- [3] 邹月,陈启愉. 基于改进 BP 小波网络的网络流量预测研究[J]. 计算机仿真, 27(6):187-193.
- [4] 陈振伟,郭拯危. 小波神经网络预测模型的仿真实现[J]. 计算机仿真, 2008, 25(6): 147-150.
- [5] 薛可,李增智,刘浏,等. 基于 ARIMA 模型的网络流量

预测[J]. 微电子学与计算机, 2004,(07):84-87.

- [6] 邹月,陈启愉. 基于改进 BP 小波网络的网络流量预测研究[J]. 计算机仿真, 27(6):187-193.

作者简介:

章治男, (1977—),副教授,硕士.研究方向为现代教育技术和计算机网络.