

基于CEGA-SVM的网络入侵检测算法

赵军

(江苏食品职业技术学院计算机应用技术系, 淮安 223004)

摘要: 针对传统遗传算法在网络入侵检测中存在分类复杂的问题, 提出结合条件熵遗传算法(CEGA)和支持向量机(SVM)的网络入侵检测算法。将入侵特征的抽取和分类模型的建立进行联合优化, 同时利用训练数据的统计特性指导入侵特征的抽取, 并对特征空间进行线性变换, 得到优化的特征子集和分类模型, 在提高分类检测率的同时降低检测时延。

关键词: 入侵检测; 遗传算法; 支持向量机

Network Intrusion Detection Algorithm Based on CEGA-SVM

ZHAO Jun

(Department of Computer Applied Technology, Jiangsu Food Science College, Huai'an 223004)

【Abstract】 Due to the complex classification problems from the traditional genetic algorithm in the process of network intrusion detection, this paper proposes the network intrusion detection algorithm combined Conditional Entropy Genetic Algorithm(CEGA) and the Support Vector Machine (SVM). To optimize jointly the invasion of feature extraction and classification model, while taking advantage of the statistical characteristics of training data to guide the invasion feature extraction, and according to the feature space linear transformation to obtain the optimal feature subset and the classification model, as improving the classification test rates, the detection latency is reduced.

【Key words】 intrusion detection; genetic algorithm; Support Vector Machine(SVM)

1 概述

网络入侵检测本质上是一个分类问题: 区分待检测数据为攻击或正常。解决这个分类问题, 主要应考虑入侵特征的选取和分类模型的建立^[1-3]。文献[4-5]采用遗传算法进行入侵特征的抽取。文献[5]提出一个基于机器学习的网络入侵检测框架, 将遗传算法用于特征抽取, 采用支持向量机在抽取后的入侵特征上建立检测模型。文献[6]提出将特征抽取和分类模型的建立相结合的算法, 在遗传个体的设计中将支持向量机的参数纳入, 在得到优化特征向量的同时也得到了分类模型的优化参数。然而这种算法只是简单地将分类错误率作为适应度函数, 没有考虑到对分类性能会产生重要影响的其他因素。

本文提出结合条件熵遗传算法(Conditional Entropy Genetic Algorithm, CEGA)和支持向量机(Support Vector Machine, SVM)的网络入侵检测技术, 在提高分类检测率的同时降低检测时延。

2 基于支持向量机的网络入侵检测

给定 m 个 k 维样本分别为 $(x_1, y_1), (x_2, y_2), \dots, (x_m, y_m)$, $x \in R^k, y \in \{-1, 1\}$ 。这里要寻找一个最佳超平面:

$$w \times x + b = 0 \quad (1)$$

能将样本分为 2 类, 则构造最佳超平面等价于找到一个向量 w_0 和一个常数 b_0 , 使得向量 w_0 具有最小的范数。考虑到可能存在被错误分类的样本, 引入松弛变量 $\xi_i \geq 0$, $i=1, 2, \dots, m$, 问题就转化为最小化式(2), 并且满足约束条件式(3):

$$\frac{1}{2} w \times w + C \sum_{i=1}^m \xi_i \quad (2)$$

$$y_i [(x_i \times w) + b] \geq 1 - \xi_i \quad (3)$$

其中, C 是一个正常数, 代表对错分的惩罚因子。

网络入侵检测模型的学习问题与一般的学习问题相比存在其特殊性。训练样本的获取十分不易, 在实际运行情况下, 某些特殊攻击样本的训练数据可能会非常少, 因此, 入侵检测问题是典型的小样本问题。与传统的统计学相比, SVM 更注重来自样本本身的信息而非产生样本的规律, 对样本的充分性要求不高, 适合于小样本的学习问题, 十分适合网络入侵检测。

3 基于条件熵遗传算法的特征抽取技术

3.1 遗传算子设计

选择操作采用轮盘赌的选择方法, 同时保留种群中的最优个体直接进入下一代的进化中。

交叉操作对实数值采用非均匀算术交叉。假设对 A, B 这 2 个个体进行交叉, 交叉后产生的个体为

$$A' = aB + (1-a)A \quad (4)$$

$$B' = aA + (1-a)B \quad (5)$$

对于二进制数值则采用单点交叉。

变异操作对于实数值采用均匀变异, 对于二进制值则采用随机变异。

在种群演化的过程中需要保留优秀个体的模式, 增强较差个体的变异能力, 才能使算法跳出局部最优解, 克服早熟的缺点。本文结合最小化适应度函数的条件, 对交叉和变异

作者简介: 赵军(1971-), 女, 讲师、工程师、硕士, 主研方向: 网络安全, 软件工程

收稿日期: 2009-07-15 **E-mail:** hyzhaojun@126.com

概率做如下改进:

(1)对于适应度函数值较大的个体,增加其交叉率和变异率。

(2)当种群中的大部分个体拥有相近的适应度且平均适应度与最小适应度接近时,说明此时群体可能收敛到局部最优解,此时应该增加大多数个体的交叉率和变异率,以跳出局部最优。

(3)改进1在进化后期,交叉率应逐渐减少,以利于保留最优个体,但变异率应逐渐提高,以跳出局部最优解。

按照以上思路,本文提出以下公式:

交叉率:

$$p_c = \begin{cases} \max[\frac{p_{cmax} + p_{cmin}}{2} + \frac{p_{cmax} - p_{cmin}}{2} F(\frac{f - f_{min}}{f_{avg} - f_{min}}) - \beta_1 \frac{t}{T}, 0] & f \leq f_{avg} \\ \max[p_{cmax} - \beta_1 \frac{t}{T}, 0] & f > f_{avg} \end{cases} \quad (6)$$

变异率:

$$p_m = \begin{cases} \min[\frac{p_{mmax} + p_{mmin}}{2} + \frac{p_{mmax} - p_{mmin}}{2} F(\frac{f - f_{min}}{f_{avg} - f_{min}}) - \beta_2 \frac{t}{T}, 0] & f \leq f_{avg} \\ \min[p_{mmax} - \beta_2 \frac{t}{T}, 0] & f > f_{avg} \end{cases} \quad (7)$$

其中, $F(x) = e^x / (e^x + 1)$; p_{cmax} 和 p_{cmin} 是交叉率的最大值和最小值; p_{mmax} 和 p_{mmin} 是变异率的最大值和最小值; t 是当前的进化代数; f 是当前个体的适应度; T 是进化的最大代数; β_1 和 β_2 是进化的代数对交叉率和变异率的影响权重,在本文中均取为 1。

按照上述公式计算得到的个体交叉率如果出现小于 0 的情况,则该个体的交叉率直接赋值为 0,个体变异率如果出现大于 1 的情况,该个体的变异率直接赋值为 1,以保证个体的交叉率和变异率在 0 和 1 之间。

由于 $F(x)$ 是增函数,因此适应度较大个体的交叉和变异概率将高于适应度小的个体,同时对于那些适应度大于平均适应度的个体,其交叉和变异概率将大于那些适应度小于平均适应度的个体,由此满足改进(1)。当种群中的大部分个体拥有相近的适应度且平均适应度与最小适应度接近时,即 $f_{avg} \rightarrow f_{min} \rightarrow 0$,此时交叉率和变异率很大,满足改进(2)。当逐渐增大时,交叉率减小,变异率增大,满足改进(3)。

3.2 算法描述

基于 CEGA-SVM 的检测算法分为训练阶段和检测阶段:

(1)训练阶段

1)数据预处理。对符号型字段编码为数值型数据,同时对所有数据进行归一化处理。设输入数据为 (d_1, d_2, \dots, d_k) , 平均值为

$$\bar{d} = \frac{1}{k} \sum_{i=1}^k d_i$$

标准差为

$$\sigma(d) = \sqrt{\frac{1}{k} \sum_{i=1}^k (d_i - \bar{d})^2}$$

则归一化后的值为

$$d'_i = \frac{d_i - \bar{d}}{\sigma(d)}$$

2)随机生成初始种群。

3)由个体的基因位确定所选择的特征、权重以及 SVM 训

练模型参数,根据适应度函数计算每个个体的适应度函数值,计算交叉率和变异率。

4)对被选中的 2 个个体进行交叉操作,产生后代个体。对被选中的个体进行变异操作。根据轮盘赌选择法按照个体的适应度函数值大小对个体进行选择操作,并保留种群中的最优个体直接进入下一代种群。由此产生新的种群。

5)重复执行 3),直到满足适应度要求或进化到最大代数,选择当前种群的最优个体作为最优解。

(2)检测阶段

根据选择的最优特征子集及其权重和优化参数建立检测模型,对待分类个体进行判断。

4 仿真与分析

4.1 仿真数据集

实验数据来源于 KDD CUP99 数据集, KDD CUP99 数据集分为训练集和测试集 2 个部分,包含了监听到的大量网络连接信息。每条连接信息包含 41 维特征,包括基本特征集、内容特征集、流量特征集和主机流量特征集。训练数据集中的每个网络连接都被标记为正常或攻击,可能的取值包括 Normal, Probe, Dos, U2R 和 R2L 5 种类型。

在仿真中,将实验数据集分为 4 个部分: Probe 数据集, Dos 数据集, R2L 数据集和 U2R 数据集。各个数据集的样本数量如表 1 所示,由于在 KDD CUP99 数据集中的 U2R 样本较少,因此仿真中使用的数据集训练样本和检测样本数量也比较少。训练集在训练阶段使用,从 KDD CUP99 数据集的训练集中随机抽取,测试集在检测阶段使用,用于评估检测模型的性能,从 KDD CUP99 数据集的测试集中随机抽取。

表 1 仿真数据集的样本数量

数据集	训练集		测试集	
	正常样本	攻击样本	正常样本	攻击样本
Probe	600	400	600	400
DoS	600	400	600	400
R2L	600	400	600	400
U2R	90	60	90	60

CEGA 的参数设置如表 2 所示,其中,交叉率和变异率的取值范围根据经验数据设定。

表 2 参数设置

参数名	参数值
种群大小	100
个体维数	84
进化最大代数	500
交叉率取值范围	{0.4, 0.8}
变异率取值范围	{0.08, 0.2}
C_1, C_2	{0.5, 0.5}

4.2 仿真结果

实验在 Matlab 7.0 环境中运行。通过对各类数据集(Probe, Dos, U2R, R2L)的测试集进行实验,CEGA-SVM 的实验结果如表 3 所示。由仿真结果可以看出,对特征进行维数约减和空间变换后,不仅入侵特征的数量基本减少了一半,而且正确检测率仍然取得了满意的结果。

表 3 CEGA-SVM 实验结果

数据集	特征数	SVM 模型参数	正确检测率/(%)
Probe	17	$C=98.7, \epsilon=2.16$	98.5
DoS	19	$C=305.3, \epsilon=0.13$	98.4
U2R	19	$C=316.9, \epsilon=0.54$	99.3
R2L	23	$C=573.2, \epsilon=0.19$	97.9

(下转第 171 页)