DOI:10.16652/j.issn.1004-373x.2018.03.029

# 基于机器学习算法的网络入侵检测

张　夏
（宜春学院，江西 宜春　336000）

**摘　要：** 网络入侵的频率越来越高，严重危害了网络安全。为了获得高正确率的网络入侵检测结果，针对当前网络入侵检测模型的局限性，提出基于机器学习算法的网络入侵检测模型，通过机器学习算法中性能优异的支持向量机构建“一对一”的网络入侵检测分类器，采用当前标准网络入侵检测数据库对模型的有效性进行验证，网络入侵检测正确率高达95%以上，检测误差远远低于实际应用范围，可以应用于实际的网络安全管理中。

**关键词：** 网络安全；入侵行为；机器学习算法；入侵检测；分类器；检测误差
**中图分类号：** TN915.08-34　　**文献标识码：** A　　**文章编号：** 1004-373X(2018)03-0124-04

## Network intrusion detection based on machine learning algorithm

ZHANG Xia
(Yichun University, Yichun 336000, China)

**Abstract:** The frequent network intrusion endangers the network security seriously. In order to obtain the network intrusion detection results with high accuracy, a network intrusion detection model based on machine learning algorithm is proposed for the limitations of the current network intrusion detection model. The support vector machine of machine learning algorithm is used to construct the one-to-one network intrusion detection classifier. The standard network intrusion detection database is used to verify the effectiveness of the model with experiment. The network intrusion detection rate is higher than 95%, the detection error is far below the actual application range. The model can be applied to the practical network security management.

**Keywords:** network security; intrusion behavior; machine learning algorithm; intrusion detection; classifier; detection error

## 0　引　言

随着网络应用的不断推广，网络安全问题受到了人们的高度关注。传统网络安全防范技术主要有数据加密、杀毒软件等，这些都是被动防范方式，无法抵挡外来的入侵行为，这样网络安全性就无法得到有效保护[1]。主动网络安全防范技术主要为入侵检测，可以对网络安全状态进行实时检测，发现一些非法的入侵行为，网络入侵成为当前研究的重点[2-3]。

当前对网络入侵检测的研究已经很深入，出现了许多性能较优的网络入侵检测模型。当前网络入侵检测模型大致可以分为误用检测和异常检测两大类。误用检测是最原始的入侵检测技术，其构建一种网络入侵检测的数据库，将待检测行为与数据库中的入侵行为进行匹配，如果匹配就将其划分到相应的入侵类别中，反之就是正常行为[4-5]。在实际应用中，误用检测模型只能检测到已经存在的入侵行为，无法检测到一些新的入侵行为，因此，当有新的入侵行为时，该模型就无能为力了，实际应用价值较低[5]。相对于误用检测技术，异常检测技术属于模式识别，通过一定的规则对入侵行为进行分析，可以检测到一些新的、从来没有出现过的入侵行为，实际应用价值相对较高，成为当前网络安全领域研究的一个重要方向[6-8]。在网络入侵的异常检测过程中，入侵行为的分类器选择十分关键，当前主要有神经网络进行网络入侵行为分类器的构建，而神经网络是一种基于大数据理论的建模方法，要求训练样本足够多，这就增加了网络入侵检测的成本，同时网络入侵实际是一种小样本，难以满足大样本的要求，因此，神经网络的网络入侵检测结果不太稳定，检测正确率时高时低，检测结果不太可信[9]。近年来，随着机器学习理论研究的不断深入，出现了一种新型建模技术——支持向量机，相对神经网络，支持向量机对训练样本数量要求没有那么高，而且学习性能也不比神经网络差，为此有学者将其引入到网络入侵检测的应用中[10]。在基于支持向量机的网络入侵检测建模过程中，存在以下难题：支持向量机参数的确定，当前对于参数确定问题，有学者采用梯度下降算

收稿日期：2017-04-06　　修回日期：2017-06-12

法、遗传算法进行寻优得到,但是梯度下降算法的寻优时间长,影响网络入侵检测的效率;遗传算法的遗传算子设置没有统一的理论指导,易获得局部最优的参数值,影响网络入侵的检测结果[11]。

为了获得高正确率的网络入侵检测结果,针对当前网络入侵检测模型的局限性,提出基于蚁群算法确定支持向量机参数的网络入侵检测模型,通过机器学习算法——支持向量机构建"一对多"的网络入侵检测分类器,采用蚁群算法确定最优参数,采用当前标准网络入侵检测数据库对模型的有效性进行测试,网络入侵检测正确率高达95%以上,检测误差远远低于实际应用范围。

## 1　相关理论

### 1.1　支持向量机

支持向量机是由Vapnik等提出的一种性能优异、专门针对小样本的机器学习算法,与神经网络的工作原理不同,其根据结构风险最小化原理进行建模,是一种二分类算法,通过找到一个最优平面,将全部训练样本划分为两类:一类位于平面上方;另一类位于平面下方。同时使样本尽可能远离最优平面,处于最优平面上的样本称之为支持向量,其工作原理如图1所示。

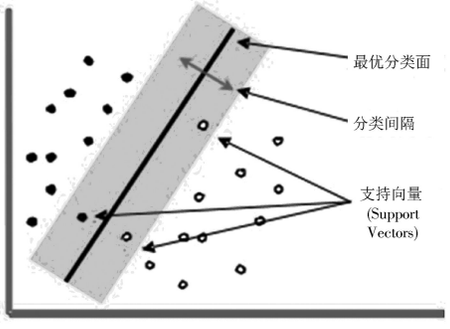


**图1　最优分类平面的示意图**

**Fig. 1　Schematic diagram of optimal classification plane**

对于含有 $n$ 个样本的集合 $\{(x_1,y_1),\cdots,(x_i,y_i),\cdots,(x_n,y_n)\}$,采用函数 $\varphi(x)$ 对样本进行映射,然后在映射空间进行样本的分类,则有:

$$f(x)=\mathrm{sgn}(w\cdot\varphi(x)+b) \tag{1}$$

式中:$w$ 为权值;$b$ 为阈值。

要找到最优分类平面,必须找到最优 $w$ 与 $b$ 值,而对于式(1)进行直接求解,得到最优 $w$ 与 $b$ 值十分困难,为此基于结构风险最小化原理,设置如下约束条件:

$$y_i\cdot\left(w\cdot\varphi(x_i)+b\right)\geqslant 1 \tag{2}$$

为了加快建模速度,采用松弛变量 $\xi_i$ 对分类精度和分类误差进行折中操作,这样最优分类平面可以转变为如下形式:

$$\min\frac{1}{2}w\cdot w+C\sum_{i=1}^{n}\xi_i \tag{3}$$

相应的约束条件为:

$$y_i(w\cdot x_i+b)\geqslant 1-\xi_i,\quad \xi_i\geqslant 0,\ i=1,2,\cdots,n \tag{4}$$

式中 $C$ 表示对误差的惩罚程度。

引入Lagrange乘子 $\alpha_i>0$,得到式(4)的对偶形式:

$$\min\frac{1}{2}\sum_{i,j=1}^{n}\alpha_i\alpha_j y_i y_j(\varphi(x_i)\cdot\varphi(x_j))+\sum_{i=1}^{n}a_i \tag{5}$$

且有如下约束条件:

$$\sum_{i=1}^{n}\alpha_i y_i=0,\quad 0\leqslant\alpha_i\leqslant C \tag{6}$$

针对非线性分类问题,引入核函数 $k(x_i,x_j)$,则可以得到:

$$\min\frac{1}{2}\sum_{i,j=1}^{n}\alpha_i\alpha_j y_i y_j k(x_i,x_j)+\sum_{i=1}^{n}a_i \tag{7}$$

式中 $k(x_i,x_j)=\varphi(x_i)\cdot\varphi(x_j)$。

支持向量机的最优分类平面为:

$$f(x)=\mathrm{sgn}\left(\sum_{i,j=1}^{n}\alpha_i y_i k(x_i,x)+b\right) \tag{8}$$

选择径向基函数,其为:

$$k(x,x_j)=\exp\left(-\frac{\left\|x-x_j\right\|}{2\sigma^2}\right) \tag{9}$$

式中 $\sigma$ 表示核宽度参数。

### 1.2　参数对网络入侵检测的影响分析

对支持向量机的工作原理进行分析可以发现,参数 $C$ 和 $\sigma$ 对其学习性能的影响十分重要。选择一个训练样本,分析不同参数条件下,网络入侵检测的正确率,结果如表1所示。对表1进行分析可知,即使环境和数据均相同,不同参数的入侵检测正确率差别仍然很大,因此需要选择参数 $C$ 和 $\sigma$ 的最优值。

**表1　参数 $C$ 和 $\sigma$ 对支持向量机学习性能的影响**

**Table 1　Influence of parameter $C$ and $\sigma$ on support vector machine learning performance**

| $C$ | $\sigma$ | 入侵检测正确率/% |
|---|---|---|
| 10 | 0.01 | 62.74 |
| 50 | 0.1 | 98.53 |
| 100 | 1 | 72.67 |
| 500 | 10 | 78.20 |
| 1 000 | 100 | 95.74 |
| 5 000 | 1 000 | 67.49 |
| 10 000 | 2 000 | 77.40 |

### 1.3　蚁群算法

蚁群算法是一种较常用的搜索优化算法。蚁群在觅食过程中,在路径上遗留下信息素,其他蚂蚁通过信

息素进行爬行路径识别，信息素浓度越高，蚂蚁经过该路径的数量就越多，其他蚂蚁选择该条路径的概率就越高，基本工作原理如图2所示。

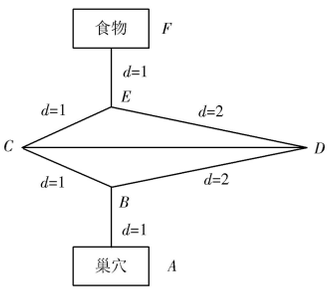


**图2 蚁群算法的工作原理**

**Fig. 2 Working principle of ant colony algorithm**

设蚂蚁数为 $m$，那么就可以得到如下计算公式：

$$m=\sum_{i=1}^{n}b_i(t) \tag{10}$$

式中 $b_i(t)$ 表示节点 $i$ 上的蚂蚁数。

在 $t$ 时刻，节点 $i$ 和 $j$ 的路径 $(i,j)$ 残留信息素浓度为：

$$\Gamma=\left\{\tau_{ij}(t)\middle|c_i,c_j\subset C\right\} \tag{11}$$

式中 $\tau_{ij}(t)$ 为 $(i,j)$ 的信息素浓度。

蚁群算法的初始工作阶段，$\tau_{ij}(0)=0$，蚂蚁选择下一个节点的转移概率 $p_{ij}^k(t)$ 为：

$$p_{ij}^k(t)=\begin{cases}\dfrac{\tau_{ij}^{\alpha}\eta_{ij}^{\beta}}{k\tau_{is}^{\alpha}\eta_{is}^{\beta}(t)}, & s\in\text{allowed},j\in\text{allowed}_k\\ 0, & \text{otherwise}\end{cases} \tag{12}$$

式中：$\eta_{ij}$ 为节点 $i$ 转移到 $j$ 的局部启发信息；$\text{allowed}_k$ 为未访问的节点集合；$\alpha$ 和 $\beta$ 为权重参数。

经过一段时间后，蚁群完成一次路径搜索，需要更新路径上的信息素，具体为：

$$\tau_{ij}(t+n)=(1-\rho)\times\tau_{ij}(t)+\Delta\tau_{ij}(t) \tag{13}$$

$$\Delta\tau_{ij}(t)=\sum_{k=1}^{m}\Delta\tau_{ij}^k(t) \tag{14}$$

式中：$\rho$ 为信息素的挥发度；$\Delta\tau_{ij}(t)$ 为路径 $(i,j)$ 上的信息素增量；$\tau_{ij}^k(t)$ 为信息素之和，其表达式为：

$$\tau_{ij}^k=\begin{cases}\dfrac{Q}{L_k}, & \text{蚂蚁}k\text{在本次循环中经过}(i,j)\\ 0, & \text{其他}\end{cases} \tag{15}$$

式中：$Q$ 为常量；$L_k$ 为本次循环的总时间。

## 2 网络入侵模型的构建

在网络入侵检测中，LSSVM参数优化问题可以采用下式进行表示：

$$\max P(C,\sigma)$$
$$\text{s.t.}\quad\begin{cases}C\in[C_{\min},C_{\max}]\\ \sigma\in[\sigma_{\min},\sigma_{\max}]\end{cases} \tag{16}$$

网络入侵检测的步骤如下：

Step1：对网络状态信息进行收集，提取网络入侵检测的特征，并对特征进行如下处理：

$$x_1=(x-x_{\min})/(x_{\max}-x_{\min}) \tag{17}$$

式中 $x_{\max}$ 和 $x_{\min}$ 分别为最大和最小值。

Step2：将支持向量机参数 $(C,\sigma)$ 看作蚁群爬行的一条路径，根据每一组参数对网络入侵检测训练样本进行建模，得到不同的检测正确率。

Step3：通过蚁群的信息素更新操作和节点转移，实现路径爬行，最后通过路径寻优找到最优的参数 $(C,\sigma)$ 组合。

Step4：根据最优的参数 $(C,\sigma)$ 组合建立最优的网络入侵检测模型。

由于支持向量机针对两个类别的分类问题，而网络入侵行为有很多种类型，如：拒绝服务攻击、未授权远程访问攻击、端口扫描攻击等。本文采用一种“一对一”的方式构建多分类器，如图3所示。

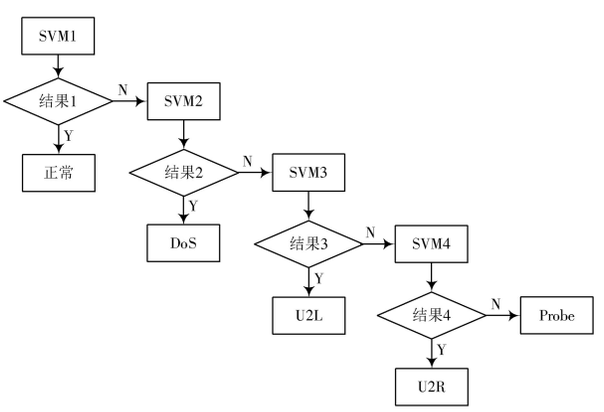


**图3 网络入侵检测的分类器结构**

**Fig. 3 Classifier structure fo network intrusion detection**

## 3 实验结果与分析

选择省理工学院的KDD Cup的网络入侵检测数据集作为测试对象，包括4种网络入侵行为：DoS，Probe，U2R和R2L。由于该数据集的规模十分庞大，为此，从中随机选取10%的数据进行具体实验。为了使本文模型（ACO-SVM）实验结果具有说服力，采用BP神经网络（BPNN）、遗传算法优化SVM（GA-SVM）的网络入侵检测模型作为对比模型，采用如下指标作为实验结果评价标准：

$$\text{正确率}=\frac{\text{正确检测样本数}}{\text{样本总数}}\times100\% \tag{18}$$

仿真实验结果如图4所示。从图4可知，在所有模型

中,ACO-SVM的网络入侵检测正确率最高,其次为GA-SVM,网络入侵检测正确率最低者为BPNN,同时误报率也最低,这表明ACO-SVM可以比较精确地实现网络入侵行为的识别,获得比较理想的检测结果。同时从图4b)可以看出,ACO-SVM网络入侵检测用时最少,可以满足网络入侵检测的效率要求,网络入侵检测结果的优越性十分明显。

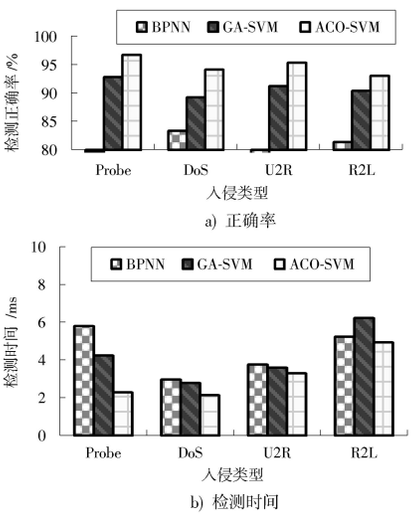


**图4　网络入侵检测结果对比**

**Fig. 4　Comparison of network intrusion detection results**

## 4　结　语

网络入侵检测的建模是一种重要网络安全防范技术,当前网络入侵检测模型无法准确刻画入侵行为,导致网络入侵检测不理想,为此设计了基于机器学习算法的网络入侵检测模型,通过支持向量机对网络入侵检测特征和网络入侵行为之间的映射关系进行拟合,建立反应两者关系的网络入侵检测模型。实验结果表明,该模型不仅可以精确地对网络入侵行为进行识别,而且检测的速度相当快,获得了比其他模型更优的网络入侵检测结果,具有广泛的应用前景。

## 参　考　文　献

[1] DENNING D E. An intrusion detection model [J]. IEEE transactions on software engineering, 2010, 13(2): 222-232.

[2] SUNG A H. Identify important features for intrusion detection using support vector machines and neural networks [C]// Proceedings of 2003 IEEE Symposium on Application and the Internet. Orlando: IEEE, 2013: 209-217.

[3] 李响.基于经验模态分解的局域网络入侵检测算法[J].西南师范大学学报(自然科学版),2016,41(8):132-137.
LI Xiang. Local network intrusion detection algorithm based on empirical mode decomposition [J]. Journal of Southwestern Normal University (natural science edition), 2016, 41(8): 132-137.

[4] 沈夏炯,王龙,韩道军.人工蜂群优化的BP神经网络在入侵检测中的应用[J].计算机工程,2016,42(2):190-194.
SHEN Xiajiong, WANG Long, HAN Daojun. BP neural network artificial bee colony optimization in the application of intrusion detection [J]. Computer engineering, 2016, 42(2): 190-194.

[5] 魏旻,王一帆,李玉,等.基于WIA-PA网络的周界入侵检测系统设计与实现[J].重庆邮电大学学报(自然科学版),2013,25(2):148-153.
WEI Min, WANG Yifan, LI Yu, et al. WIA-PA network based perimeter intrusion detection system design and implementation [J]. Journal of Chongqing University of Post and Telecommunications (natural science edition), 2013, 25(2): 148-153.

[6] VILAPLANA V, MARQUES F, SALEMBIER P. Binary partition trees for object detection [J]. IEEE transactions on image processing, 2010, 17(11): 2201-2216.

[7] HONG J, SU M Y, CHEN Y H, et a1. A novel intrusion detection system based on hierarchical clustering and support vector machines [J]. Expert systems with applications, 2011(38): 306-313.

[8] MUNI D P, PAL N R, DAS J. Genetic programming for simultaneous feature selection and classifier design [J]. IEEE transactions on systems, man, and cybernetics: part B, 2009, 36(1): 106-117.

[9] KENNEDY J, EBERHART R C. Particle swarm optimization [C]// Proceedings of 2005 IEEE International Conference on Neural Networks. Perth: IEEE, 2005: 1942-1948.

[10] 杨宏宇,赵明瑞,谢丽霞.基于自适应进化神经网络算法的入侵检测[J].计算机工程与科学,2014,36(8):1469-1475.
YANG Hongyu, ZHAO Mingrui, XIE Lixia. Intrusion detection based on adaptive evolutionary neural network algorithm [J]. Computer engineering and science, 2014, 36(8): 1469-1475.

[11] 江峰,王春平,晋惠芬.基于相对决策熵的决策树算法及其在入侵检测中的应用[J].计算机科学,2012,39(4):223-226.
JIANG Feng, WANG Chunping, JIN Huifen. Decision tree algorithm based on relative decision entropy and its application in intrusion detection [J]. Computer science, 2012, 39(4): 223-226.

[12] 龚俭,王卓然,苏琪,等.面向网络安全事件的入侵检测与取证分析[J].华中科技大学学报(自然科学版),2016,44(11):30-33.
GONG Jian, WANG Zhuoran, SU Qi, et al. Intrusion detection and forensics analysis for network security incidents [J]. Journal of Huazhong University of Science and Technology (natural science edition), 2016, 44(11): 30-33.

**作者简介:**张　夏(1985—),女,河北保定人,硕士,讲师。主要研究方向为计算机网络、人工智能。