以机器学习为内核的网络入侵检测模型研究

顾凌云

（上海冰鉴信息科技有限公司，上海 200120）

摘 要：网络技术实现了万物互联，为人们的生活、学习和工作等都带来了极大便利，然而随着网络环境的越来越复杂，网络系统安全风险也日益突出，尤其是网络入侵对网络安全的威胁巨大，加强网络入侵检测是保证网络安全的第一关口，有着重要意义。本文基于支持向量机、蚁群算法等原理构建基于机器学习算法的网络入侵检测模型，该模型经标准网络入侵检测数据库的验证，其检测准确率高达95%，可很好地提升网络安全。

关键词：网络安全 机器学习 入侵检测模型 支持向量机 蚁群算法

中图分类号：TN711 **文献标识码：**A **文章编号：**1003-9082（2022）06-0004-03

引言

网络安全维护是一项非常复杂的系统工程，仅靠传统单一的防御手段和低级别的技术手段难以达到理想效果，借助更高水平的网络入侵检测技术解决此类问题的趋势不可阻挡。笔者在此解析了网络入侵及检测、机器学习的概念，分析了支持向量机、蚁群算法等原理，然后构建基于机器学习的网络入侵模型，最后对模型检测质量予以实验验证。

一、基本概念解析

1.网络入侵及检测技术

网络入侵是以网络为渠道的非法侵害行为。当前较为常见的网络入侵路径主要包括经协议缺陷入侵、经系统漏洞入侵和以病毒程序寄生系统。所谓的网络入侵检测，是专门针对网络入侵行为做出的反应，即监测、发现存在的各种已知和未知的网络访问异常并对其做出警示，其入侵检测系统（IDS）主要由数据包嗅探、数据预处理器、网络检测引擎、报警和日志模块四部分构成（见图1）。数据包嗅探主要负责监听网络数据包，对网络行为进行数据采集；数据预处理器对网络数据包内容进行初步提炼，发现原始数据中的“异常现象”，形成可供分析的结构化的数据内容；检测引擎依据预先设置的相关规则来检查数据包，一旦发现异常立即反馈给报警模块；报警和日志模块就是存储引擎提供的异常信号并发出警示[1]。

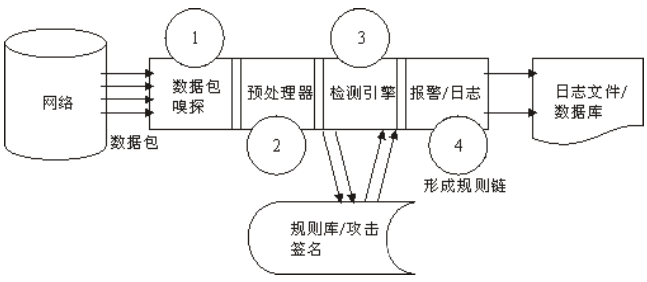

图1 网络入侵检测系统结构图

2.机器学习

机器学习（Machine Learning）是一种基于大数据环境的计算机模拟人类学习行为的过程。它可以在一定程序设定下进行类人学习活动并且具备自动重组已有知识结构实现功能升级的能力。机器学习涉及众多算法，任务和学习理论，从任务类型来看，机器学习模型可包括回归模型、分类模型和结构化学习模型；从学习方法层面来看，机器学习可分为线性模型和非线性模型，非线性模型包括SVM、KNN等传统模型和深度学习模型；此外根据学习理论还可以将其划分为有监督学习、半监督学习、无监督学习、迁移学习和强化学习几种[2]。

二、网络入侵检测的相关原理

1.支持向量机

支持向量机（SupportVectorMachine，SVM）是一种二分类模型，它的基本模型是定义在特征空间上的间隔最大

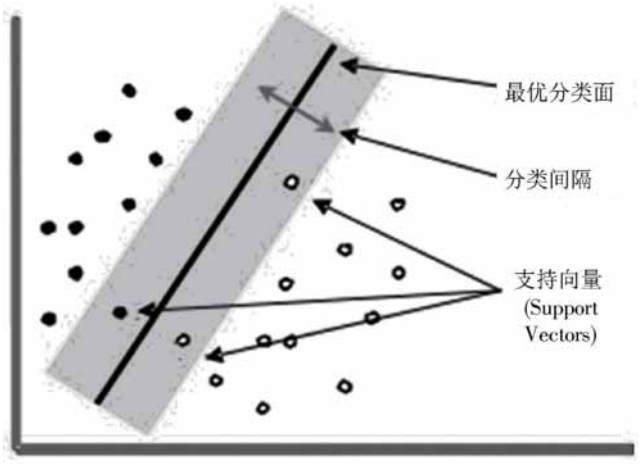

图2 最优分类平面的示意图

的线性分类器，其学习策略便是间隔最大化；SVM最基本的应用是分类，其学习算法就是求解凸二次规划的最优化算法，即求解最优的分类面[3]。SVM性能优异属于专门针对小样本的机器学习算法，其工作原理类似于神经网络，是基于结构风险最小化理论之上在特征空间构建最优分隔超平面，以此让学习器实现全局最优化，并且在整个样本空间的期望风险以某个概率满足一定上界，其原理如图2所示。

以函数ϕ（x）对具有n个样本的集合{（x1，y1），…，（xi，yi），…，}（xn，yn）进行映射处理，样本划分也在映射空间进行，以关系式表示：

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

该算式中以w作为权值，以b表示阈值。通过找到最优的w值和b值从而确立最优分类平面，但是要想以直接求解算式（1）而得出最优的w与b值并不容易，可以依据结构风险最小化原理，设置约束关系式：

$$y_i\cdot[w\cdot\varphi(x_i)+b]\geq 1$$（2）

处于快速建模考虑，可以采用松弛变量 ξ_i 来折中处理分类精度及分类误差，得到如下形式的最优分类平面：

$$\min\frac{1}{2}w\cdot w+c\sum_{i=1}^{n}\xi_i$$（3）

针对上式的越约束条件为：

$$y_i\cdot(w\cdot x_i+b)\geq 1-\xi_i$$（4）

式中 $\xi_i\geq 0,i=1,2.3.....,n$　C则代表基于测算误差的惩罚程度。通过引入Lagrange乘子 a_i>0，可以得出上式的对偶形式：

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

其约束条件设置为：$\sum_{i=1}^{n}a_i y_i=0,0\leq a_i\leq C$（6）

然后基于非线性分类相关问题，将核函数k（x_i，x_j）引入算式（5）则可以得到：

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

算式中 $K(x_i,x_j)=\varphi(x_i)\cdot\varphi(x_i)$ 并由此得到支持向量机的最优分类平面：

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

选择径向基函数，由此得到：

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

上面算式中σ代表核宽度参数值。

2.网络入侵检测中的参数影响作用

结合以上支持向量机的工作原理，经过分析我们可以发现，参数核宽度参数σ和测算误差惩罚参数C能够对机器的学习性能产生一定影响。在此选择一批训练样本，对其不同参数条件下的网络入侵检测准确度进行分析，得到如表1所示结果。

表1　参数C与参数σ对支持向量机学习性能的影响情况

c	σ	入侵检测准确率（%）
10	0.01	62.74
50	0.1	98.53
100	1	72.67
500	10	78.20

由对表1数据的分析可以发现，即便在相同环境和数据条件下，不同参数的入侵检测效果依然会出现较大差异，所以必须选择最优的C和σ参数值。

3.蚁群算法

蚂蚁在觅食过程中分工明确，工蚁会在觅食路线和食物附近留下具有自身独特辨识性的生物信息素，从而便于其他蚂蚁跟从，留下的信息素浓度越高则越便于蚁群识别，并将食物顺利搬入巢穴[4]。蚁群算法就是基于这种生物特征的一种非常形象的信号线索优化算法。该算法的基本工作原理详见图3所示。

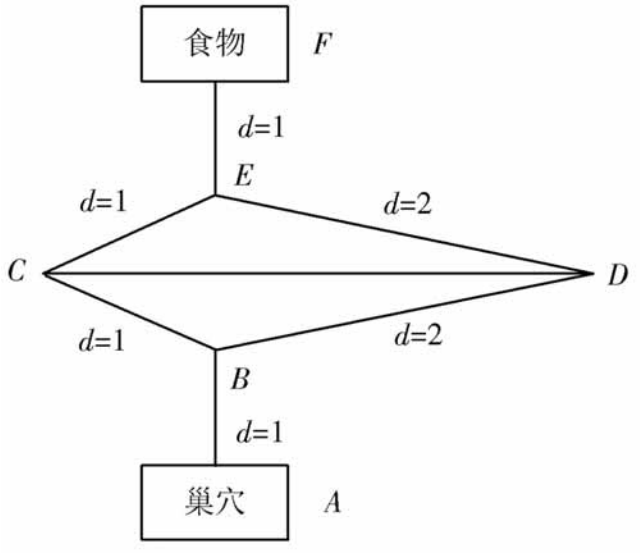

图3　蚁群算法的工作原理

如果假设蚂蚁数量为m，可以得到以下计算公式：

$$m=\sum_{i=1}^{n}b_i(t)$$（10）

算式中 $b_i(t)$ 指的是 节点i上的蚂蚁数量。那么在t时刻，

i节点和j节点上的路径（i，j）所留下的蚂蚁信息素浓度可以表示如下：

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

其中$\tau_{ij}(t)$就是（i，j）路径上的信息素浓度。

当蚁群算法初始点$\tau_{ij}(t)$（0）=0，那么蚂蚁对于下一个节点选择的转移概率$p_{ij}^{k}(t)$可以按照以下算式进行计算：

$$p_{ij}^{k}(t)=\begin{cases}\dfrac{\tau_{ij}^{a}\eta_{is}^{\beta}}{k\tau_{is}^{a}\eta_{is}^{\beta}(t)} & s\in \text{allowed}, j\in \text{allowed}\\ o, & \text{other wise}\end{cases}$$（12）

上式中：η_j表示从i节点转移到j节点的局部启发信息；allowed k代表没有访问的节点集合；α与β则代表权重参数。经过一段时间蚁群在完成一次路径搜索后开始寻找新的路径信息素，可以表示为以下关系式：

$$\tau_{ij}(t+n)=(1-\rho)\times\tau_{ij}(t)+\Delta\tau_{ij}(t)$$
$$\Delta\tau_{ij}(t)=\sum_{k=1}^{m}\Delta\tau_{ij}^{k}(t)$$（13）

其中：ρ指的是信息素的挥发度；$\Delta\tau_{ij}(t)$则指的是路径（i，j）上的信息素增量；

$\tau_{ij}^{k}(t)$就是所有的信息素之和，得到如下关系式：

$$\tau_{ij}^{k}=\begin{cases}\dfrac{Q}{L_K} & \text{蚂蚁k在本次循环本次循(i, j)}\\ 0, & \text{other wise}\end{cases}$$（14）

算式中Q是一个常量；L_K表示的是该次循环的总时间数。

三、基于机器学习的网络入侵模型构建

结合上述算法原理，采用如下算法逻辑对网络入侵检测中的参数进行优化：

$$\max p(c,\sigma)$$
$$s\cdot t\cdot\begin{cases}c & \in[c_{\min},c_{\max}]\\ \sigma & \in[\sigma_{\min},\sigma_{\max}]\end{cases}$$（15）

网络入侵检测步骤如下：

步骤一：收集有关于网络状态的所有信息并从中提取出网络入侵检测特征，然后对这些特征进行以下处理：

$$x_1=\frac{(x-x_{\min})}{(x_{\max}-x_{\min})}$$（16）

步骤二：以支持向量机中（c，σ）作为一种蚁群爬行路径，结合各组参数中的网络入侵检测训练样本构建检测模型，从而可以获得不同的检测准确率。

步骤三：根据更新操作蚁群信息素以及节点转移，以此实现路径爬行，然后按照路径最优原则找到最佳的（c，σ）参数组合。

步骤四：利用得到的最佳（c，σ）参数组合来构建最优的网络入侵检测模型。鉴于支持向量机仅有两种类型的划分，而网络入侵行为多种多样，文章中的多分类器使用“一对一”的方式来构建。

四、网络入侵检测效果验证

本次实验测试对象为KDD Cup网络入侵检测数据集，其主要的网络入侵行为：包括DoS、U2R、R2L几种。我们从该庞大数据集中随机选取其1/10数据进行实验。为测试本文构建的（ACO-SVM）模型效果，将其与BP神经网络（BPNN）和遗传算法SVM（GA-SVM）网络入侵检测模型进行对比，并进行评价：

检测正确率A=$\frac{N}{U}$×100%，其中U表示样本总数，N表示检测样本例数。检测结果显示：正确率最高的是ACO-SVM其正确率高达95%，其次是GA-SVM正确率平均90%，检测效果最差的是BPNN，正确率低于85%。由此可知ACO-SVM在准确捕捉网络入侵行为方面效果较好。另外，根据比对结果可知，检测时间方面ACO-SVM相比于另外两类检测技术其用时最少，一般为2-5ms，具有更好的检测时效性。综上所述，基于机器学习的ACO-SVM网络入侵检测模型其对入侵行为的检测质量更好，具有较大的应用价值。

结语

加强网络入侵检测是确保网络安全的重要手段，以机器学习为核心的网络入侵检测模型相比于传统网络防控技术具有独特的技术优势。通过试验证实以机器学习为主的网络入侵检测模型，不但可以更加准确地检测出多种复杂入侵行为，而且检测的效率更高，因此具有极其广阔的应用和发展前景。

参考文献

[1]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(2):190-194.

[2]诸俊.计算机网络安全入侵检测技术分析[J].电子技术与软件工程,2015(09):233.

[3]张夏.基于机器学习算法的网络入侵检测[J].现代电子技术,2018(3):124-127.

[4]魏小涛,黄厚宽,田盛丰.在线自适应网络异常检测系统模型与算法[J].计算机研究与发展,2010,47(3):485-492.