

人工蜂群优化支持向量机算法在网络安全中的应用

梁海军

(淮海工学院 信息中心, 江苏 连云港 222005)

摘 要: 针对支持向量机(SVM)性能的影响,探讨了人工蜂群算法(ABC)对 SVM 参数优化方法,建立了 SVM 参数优化模型,并将其用于网络安全中的网络入侵模型中.采用 KDD 1999 数据集进行仿真实验,验证了方法的有效性,结果表明,与遗传算法等传统优化算法相比,ABC 优化的 SVM 有效地降低运行时间,可以获得更高的网络入侵检测率.

关键词: 人工蜂群算法;支持向量机;参数优化;网络安全

中图分类号: TP18

文献标识码: A

文章编号: 1000-7180(2013)02-0095-04

Application of Support Vector Machine Optimized by Artificial Bee Colony Algorithm in Network Security

LIANG Hai-jun

(Data Center, Huaihai Institute of Technology, Lianyungang 222005, China)

Abstract: Parameters of support vector machine (SVM) is the key factor that impacts its performance. A parameter optimization method for SVM using artificial bee colony algorithm (ABC) is discussed and a parameter optimization model is established. The proposed algorithm's performance is tested by KDD 1999 data. The results show that the proposed algorithm decreased the running time efficiently and improved the accuracy of network intrusion detection.

Key words: artificial bee colony algorithm; support vector machine; parameters optimization; network security

1 引言

随着网络规模的日益扩大和开放,网络安全问题已成为网络技术发展的前沿研究课题.当前主要有防火墙、杀毒软件和网络入侵检测等网络安全技术,防火墙、杀毒软件等是被动防范措施,无法满足现代网络需求,网络入侵检测系统作为一种主动网络安全防范措施,已成为当前网络安全领域的研究热点^[1].

目前,人们对于网络入侵检测的方法有很多,主要包括贝叶斯网络、K 近邻法、神经网络等,这 3 种算法在网络入侵检测预测中取得了良好效果^[2-4].在实际应用中,网络入侵特征与入侵类别之间关系复杂,这些算法呈现不足之处.支持向量机(support vector machine, SVM)是一种针对小样本的模式识别方法,对样本数量的依赖性弱,克服神经网络存在

缺陷,在网络入侵检测领域得到了越来越广泛的应用^[5].在利用 SVM 进行网络入侵检测过程中, SVM 参数选择对入侵检测结果具有至关重要的影响.目前常用的 SVM 参数优化方法有梯度下降法、遗传算法和粒子群优化算法等^[6].实践证明, SVM 的网络入侵检测结果与参数间存在多峰值函数关系,而这些算法均会不同程度地陷入局部最优解,无法得到最优网络入侵检测正确率^[7].

为获得更合理的 SVM 参数,提高网络入侵检测性能,提出一种人工蜂群算法和支持向量机相结合的网络入侵检测模型(ABC-SVM). ABC-SVM 利用人工蜂群算法(artificial bee colony algorithm, ABC)对 SVM 参数自动搜索和确定,采用 KDD 1999 网络入侵检测数据集对 ABC-SVM 有效性进行检验.

2 支持向量机性能影响分析

设给定样本集 $\{(x_1, y_1), \dots, (x_i, y_i), \dots, (x_n, y_n)\}$, 通过非线性映射函数 $\varphi(x)$, 将原始输入空间的数据映射到高维特征空间, 并在高维特征空间中进行最优分类超平面构造^[8]:

$$f(x) = w \cdot \varphi(x) + b = 0 \quad (1)$$

式中, w 表示权值矢量, b 表示阈值.

为了使得结构风险最小化, 最优分类平面应该满足如下约束条件:

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

引入非负松弛变量 ξ_i , 来提高学习方法的分类泛化能力. 式(1)优化问题变成:

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

约束条件为:

$$y_i (w \cdot x_i + b) \geq 1 - \xi_i, \quad \xi_i \geq 0, \quad i = 1, \dots, n \quad (4)$$

式中, C 为误差惩罚因子.

通过引入 Lagrange 乘子算法, 将上述优化问题转化为对偶形式:

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

同时满足

$$\sum_{i=1}^n \alpha_i y_i = 0, \quad C \geq \alpha_i \geq 0 \quad (6)$$

式中, $\alpha_i > 0$ 对应的点称为支持向量.

对于非线性分类问题, 通过引入核函数 $k(x_i, x_j)$, 得最 SVM 分类模型为:

$$f(x) = \text{sign} \left(\sum_{i,j=1}^n \alpha_i y_i k(x_i, x) + b \right) \quad (7)$$

径向基函数(RBF)参数少(只有一个参数 σ), 选择 RBF 作为 SVM 的核函数.

为了直观说明 SVM 参数 C 和 σ 对分类结果的影响, 采用网络入侵数据进行分析, 为了简化问题, 将所有非正常网络行为归为入侵类, 它们 10 折交叉验证结果见表 1 和表 2.

表 1 $C=50, \sigma$ 值对分类结果的影响

σ	网络入侵检测正确率/%
0.1	76.49
0.5	80.65
1	90.22
10	87.20

表 2 $\sigma=0.5, C$ 值对分类结果的影响

C	网络入侵检测正确率/%
0.1	75.80
1	89.86
10	88.97
100	88.57

由表 1 和表 2 的分类结果可知, 相同数据集下, 不同 C 和 σ 得到的分类精度不同, 要获取性能优越的 SVM, 需要选取最合适的 C 和 σ , 因此 SVM 训练模型参数优化的数学模型为:

$$M = \{\sigma, C\} \quad (8)$$

由于 SVM 参数优化目标是提高网络入侵的检测正确率, 因此用 SVM 对网络入侵检测正确率(G)作为 SVM 参数优化目标函数, SVM 训练模型参数的优化问题可以描述为:

$$\max_M G(M) \quad (9)$$

其约束条件为:

$$\begin{cases} M = \{\sigma, C\} \\ \sigma > 0 \\ C > 0 \end{cases} \quad (10)$$

3 人工蜂群算法的支持向量机参数优化模型

3.1 人工蜂群算法

在人工蜂群算法(ABC)算法中, 蜜蜂被分成: 采蜜蜂、观察蜂; 侦查蜂 3 种类型. 在 ABC 算法中, 每个食物源的位置代表优化问题的一个可能的解, 每个食物源的花蜜量对应每个解的适应度. ABC 算法在求解优化问题时, 食物源的位置被抽象成解空间中的点, 蜜蜂采蜜的过程也就是搜寻最优解的过程.

首先, ABC 算法生成含有 SN 个解(食物源)的初始种群, 每个解 $X_i (1, 2, \dots, SN)$ 是一个 d 维的向量. 然后, 蜜蜂对所有的食物源进行循环搜, 采蜜蜂首先对相应的解进行一次邻域搜索, 若搜索到的解的适应度优于原来解的适应度, 将原始解替换为新解. 所有的采蜜蜂完成搜索之后, 回到蜂巢, 通过跳摇摆舞传达给观察蜂; 观察蜂根据收集到的信息按照与花蜜量(适应度)相关的概率选择一个食物源. 观察蜂选中食物源后, 并像采蜜蜂一样进行一次邻域搜索, 并保留较好的解, 不断重复搜索, 最终找到最优解.

采蜜蜂和观察蜂依据公式(11)进行食物源位置更新公式为:

$$v_{ij} = x_{ij} + r_{ij} (x_{ij} - x_{kj}) \quad (11)$$

式中, $k \in (1, 2, \dots, SN)$, $j \in (1, 2, \dots, d)$, 是随机选取下标, $r_{ij} \in [-1, 1]$ 之间一个随机数, 它控制 x_{ij} 邻域的生成范围, 随着搜索接近最优解, 邻域的范围会逐渐减小。

观察蜂选择食物源概率 P_i 计算下式为:

$$P_i = \frac{fit_i}{\sum_{j=1}^{SN} fit_j} \quad (12)$$

式中, fit_i 是第 i 个解的适应度值。

假定经过 $limit$ 次循环之后, 解没有得到改善, 表明这个解陷入局部最优, 该位置对应解就要被放弃, 与这个解相对应的采蜜蜂转变为侦察蜂, 寻找新的食物源, 即

$$x_i = x_{min} + rand(0, 1)(x_{max} - x_{min}) \quad (13)$$

3.2 人工蜂群算法对支持向量机参数优化

(1) 初始化 ABC 算法中的控制参数, 主要包括: 食物源的数量 SN , 即采蜜蜂的数量; 食物源最大循环次数 N ; 终止循环次数 N_{mc} 。

(2) 采用实数对 SVM 参数 (C, σ) 进行编码。每个解个体的编码由一个实数行向量组成, 它与 C 和 σ , 是一种可能的优化个体。

(3) 在编码空间中, 随机生成一个具有 SN 个个体的初始群体。

(4) 计算当前群体中所有个体的适应度值。由于优化 SVM 参数目的在于提高网络入侵检测正确率, 适应度函数定义为:

$$v_{obj} = \frac{1}{1 - V_{acc}} \quad (14)$$

式中, V_{acc} 为网络入侵检测正确率。

(5) 采蜜蜂根据公式(11)做邻域搜索产生新解 V_i , 并且计算其适应度值。

(6) 如果 V_i 的适应度值好于 X_i , 则用 V_i 替换 X_i , 将 V_i 作为当前最好解, 否则保留 X_i 不变。

(7) 计算 X_i 的适应度值, 并根据公式(12) 计算与 X_i 相关的概率值 P_i 。

(8) 观察蜂根据 P_i 选择食物源(解), 并根据公式(12) 进行邻域搜索产生新解 V_i , 计算其适应度值; 采用步骤(4) 方式进行替换。

(9) 判断是否有要放弃的解, 如果存在, 则侦查蜂根据公式(13) 产生一个新解 X_i 代替它。

(10) 记录迄今为止最好的解; 是否满足循环终止条件, 如满足则输出最优结果, 否则返回(13)。

(11) 把最后一代的 SN 个个体全部解码, 得到 SN 组 SVM 的 C, σ 。

(12) 根据(10) 得到 C, σ , 对 SVM 进行训练, 建立最优 SVM 模型, 输入测试样本进行检测。

4 仿真实验

4.1 数据集

为了评估模型的有效性, 采用 KDD 1999 网络入侵标准数据集进行仿真实验, 该数据集包含拒绝服务攻击(DoS)、扫描与探测(Probe)、未经授权的远程访问(R2L)和对本地超级用户的非法访问(U2R)四种攻击方式, 其它数据均为正常数据。数据见表 3。

表 3 实验样本集分布情况

入侵类型	正常样本	入侵样本
DoS	780	260
Probe	900	300
R2L	75	25
U2R	100	35

4.2 实验环境和对比模型

在 Intel 奔腾 4 2.7G CPU 双核, 内存为 1GB 的计算机上实现, SVM 软件采用台湾林智仁教授的 LIBSVM 2.86。为了使 ABC-SVM 结果更具说服力, 采用遗传算法优化 SVM 参数模型(GA-SVM) 和粒子群算法优化 SVM 参数模型(PSO-SVM) 进行对比实验。

4.3 结果与分析

ABC-SVM、GA-SVM 和 PSO-SVM 对表 4 的网络入侵数据进行训练, 得到的最优参数如表 2 所示, 然后采用最优参数对测试集进行检测, 得到结果如图 1 所示。

表 4 各模型得到的最优参数

模型	C	σ
GA-SVM	100	0.625
PSO-SVM	125	0.155
ABC-SVM	50	1.925

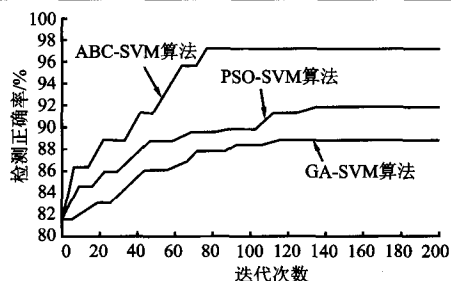


图 1 各种模型的网络入侵检测性能对比

从图 1 可知, GA-SVM 迭代 130 代左右, 网络入侵检测正确率达到最高的 88.6%; PSO-SVM

迭代 150 代左右,网络入侵检测正确率达到最高的 91.6%;而 ABC-SVM 迭代 80 代左右,网络入侵检测正确率达到最高的 96.6%,对比结果表明,ABC-SVM 的收敛速度更快,而且检测性能更优。

在现代大规模网络入侵检测中,检测速度十分关键,采用模型检测时间作为检测速度评价指标。速度检测结果如表 5 所示。

表 5 不同模型的检测速度对比

入侵类型	GA-SVM	PSO-SVM	ABC-SVM
Probe	21.49	18.66	11.98
DoS	17.63	11.37	9.62
U2R	4.17	3.43	2.44
R2L	2.27	1.60	1.16

从表 5 可知,ABC-SVM 检测速度最快,主要是由于加快了参数优化收敛速度,以更短的时间找到 SVM 参数,减少计算时间和支持向量数,从而加快了网络入侵检测速度。

5 结束语

SVM 参数对 SVM 的网络入侵检测结果有着重要影响,为克服传统参数选择方法对 SVM 性能的不利影响,提出了采用 ABC 对 SVM 参数进行优化,并应用于网络入侵检测。以网络入侵检测结果作为适应度函数,RBF 核函数作为 SVM 的核函数,利用 ABC 对惩罚因子 C 和宽度参数 σ 进行优化。采用 KDD 1999 数据集进行仿真实验,与 GA、PSO 相比,ABC 克服局部极值点缺陷,获得更高的网络入侵检测正确率。并且有效地降低参数搜索时间,加快

网络入侵检测速度。在研究过程中发现,更加有效的确定 ABC 自身参数,将会进一步提高 SVM 的网络入侵检测模型的效率和实用价值。

参考文献:

- [1] 陈秀真. 网络化系统安全态势评估的研究[J]. 西安交通大学学报, 2004(38): 404-408.
- [2] 李平均, 刘权, 李天. 基于主被动结合的网络入侵检测系统的设计与实现[J]. 微电子学与计算机, 2007, 24(7): 50-52.
- [3] 向西西, 黄宏光, 李予东. 基于 Kalman 算法的网络安全态势预测方法[J]. 计算机仿真, 2010, 27(12): 113-117.
- [4] 周荃, 王崇骏, 王王君, 等. 基于人工智能技术的网络入侵检测的若干方法[J]. 计算机应用研究, 2007, 24(5): 144-149.
- [5] 燕飞, 秦世引. 一种基于模拟退火的支持向量机超参数优化算法[J]. 航天控制, 2008, 26(5): 7-11.
- [6] 丁世飞, 齐丙娟, 谭红艳. 支持向量机理论与算法研究综述[J]. 电子科技大学学报, 2011, 40(1): 2-10.
- [7] 吴景龙, 杨淑霞, 刘承水. 基于遗传算法优化参数的支持向量机短期负荷预测方法[J]. 中南大学学报: 自然科学版, 2010, 40(1): 180-184.
- [8] 徐飞, 徐卫亚, 王珂. 基于蚁群优化最小二乘支持向量机模型的边坡稳定性分析[J]. 工程地质学报, 2009, 17(2): 253-257.

作者简介:

梁海军 男, (1977-), 工程师. 研究方向为计算机网络.