

杜鹃鸟搜索算法优化最小二乘支持 向量机的网络入侵检测模型

陈 健¹, 陈雪刚², 张家录³, 程杰仁⁴

(1 防灾科技学院 教务处, 河北 三河 065201; 2 湘南学院 计算机科学系, 湖南 郴州 423000;

3 湘南学院 数学系, 湖南 郴州 423000; 4 杭州电子科技大学 软件学院, 浙江 杭州 310018)

摘 要: 为了提高网络入侵检测率, 提出一种改进杜鹃鸟搜索(MCS)算法优化最小二乘支持向量机(LSSVM)的入侵检测模型(MCS-LSSVM). 首先将 LSSVM 的参数作为杜鹃鸟的鸟巢位置, 然后通过模拟杜鹃鸟种群寄生繁衍机制, 找到鸟巢最优位置, 并转化成 LSSVM 最优参数, 最后采用最优参数建立最优建立入侵检测模型, 并采用 KDD CUP 99 数据集进行仿真测试. 仿真结果表明, MCS-LSSVM 不仅提高了网络入侵的检测率, 而且减少了训练时间, 提高网络入侵检测效果.

关键词: 网络入侵; 杜鹃鸟算法; 最小二乘支持向量机; 参数优化

中图分类号: TP391

文献标识码: A

文章编号: 1000-7180(2013)10-0029-04

Network Intrusion Detection Model Based on Least Squares Support Vector Machine Optimized by Modify Cuckoo Search Algorithm

CHEN Jian¹, CHEN Xue-gang², ZHANG Jia-lu³, CHENG Jie-ren⁴

(1 Institute of Disaster Prevention Science and Technology, Sanhe 065201, China;

2 Department of Computer Science, Xiangnan University, Chenzhou 423000, China;

3 Department of Mathematics, Xiangnan University, Chenzhou 423000, China;

4 Institute of Software, Hangzhou Dianzi University, Hangzhou 310018, China)

Abstract: In order to improve detection rate of the network intrusion, this paper proposes a intrusion detection model (MCS-LSSVM) which the least square support vector machine(LSSVM) is optimized by improved cuckoo search(MCS) algorithm. Firstly, the parameters of LSSVM are taken as a nest location of cuckoo, and then the cuckoo species parasitic generation mechanism is used to find the optimal nest and transformed into LSSVM optimal parameters, finally, the optimal parameters is used to establish intrusion detection models. The simulation results show that the proposed model not only improves the detection rate of network intrusion, and shorten the training time to improve the network intrusion detection effect.

Key words: network intrusion; cuckoo search algorithm; least square support vector machine; parameters optimization

1 引言

随着网络规模的日益扩大和开放, 网络攻击在数量和危害程度上均呈上升态势, 传统网络安全防范措施无法满足现代网络安全需要, 网络入侵检测系统作为一种主动网络安全防范措施, 成为网络安

全领域的研究热点^[1].

入侵检测通常分为误用入侵检测(misuse intrusion detection, MID)和异常入侵检测(anomaly intrusion detection, AID)两种^[2]. 误用入侵检测技术检测率高, 但不能发现变异或新的入侵行为, 异常入侵检测技术可以发现未知攻击入侵, 因此, 当前主

收稿日期: 2012-12-01; 修回日期: 2013-01-22

基金项目: 防灾科技学院重点资助课题(2012A17)

要针对异常检测进行建模^[3]. 网络入侵检测实质是一个多分类问题, 主要包括特征选择和分类器设计, 本文主要研究分类器设计. 最小二乘支持向量机 (least square support vector machine, LSSVM) 是一种专门针对小样本、高维、非线性问题的机器学习算法, 较好地克服了神经网络的过拟合、“维数灾”等问题, 泛化能力优异, 在网络入侵检测应用最为广泛^[4]. 在实际应用中, 最典型难题是 LSSVM 参数优化. 为了解决此难题, 学者们提出采用遗传算法、粒子群算法、模拟退火算法、蚁群算法等对 LSSVM 参数进行优化, 但这些算法均存在各自不足, 影响 LSSVM 的网络入侵检测率. 杜鹃鸟搜索 (cuckoo search, CS) 算法是一种群体智能优化方法, 结合了鸟类及果蝇特殊的利维飞行模式进行搜索, 并增加了群体之间的信息交流, 加速收敛效率^[5].

为了提高 LSSVM 的网络入侵检测率, 将杜鹃鸟搜索算法引入到 LSSVM 参数优化中, 并提出一种改进杜鹃鸟搜索算法 (modify cuckoo search algorithm, MCS) 优化 LSSVM 的网络检测模型 (MCS-SLSVM), 并通过仿真实验对算法有效性和可行性进行验证.

2 杜鹃鸟算法和最小二乘支持向量机

2.1 杜鹃鸟算法

2009 年, Yang 与 Deb 等提出了杜鹃鸟搜索算法, 其基于杜鹃鸟种群 Pa 寄生繁衍策略, 为了模拟杜鹃鸟寻巢的行为, CS 算法设定了三个规则, 具体如下:

(1) 杜鹃鸟一次下一个蛋, 代表问题的一种解决方案, 并随机放在一个鸟巢中进行孵化.

(2) 一部分鸟巢放着优质蛋, 即好的解决方案, 这些鸟巢将被保留到下一代.

(3) 可利用鸟巢的数量是固定的, 鸟窝的主人能发现一个外来鸟蛋的概率为 $P_a \in (0, 1)$, $x_i^{(t)}$ 表示第 i 个鸟巢在第 t 代的鸟巢位置, 那么杜鹃鸟寻巢的路径和位置更新公式为

$$x_i^{(t+1)} = x_i^{(t)} + \partial \oplus L(\lambda), i = 1, 2, \dots, n \quad (1)$$

式中, ∂ 表示步长控制量; $\oplus$ 表示点对点乘法, $L(\lambda)$ 表示随机搜索路径^[6].

在基本 CS 算法中, 步长产生具有随机性, 缺乏自适应性, 无法保证快速收敛^[7]. 为解决全局寻优能力和精度间的关系, 在基本 CS 算法基础上, 根据不同阶段的搜索结果, 对步长大小进行自适应动态调整, 离最优鸟巢位置的自适应步长调整策略为

$$step_i = step_{\min} + (step_{\max} - step_{\min})d_i \quad (2)$$

式中, $step_{\max}$ 和 $step_{\min}$ 分别表示最大和最小步长.

d_i 定义如下:

$$d_i = \frac{\|n_i - n_{\text{best}}\|}{d_{\max}} \quad (3)$$

式中, n_{best} 表示当前鸟巢位置的最佳状态, n_i 表示第 i 个鸟巢位置, $d_{\max}$ 表示最优位置与其他鸟巢间的最大距离.

2.2 最小二乘支持向量机

LSSVM 将支持向量机的不等式约束改成等约束, 用二次损失函数代替支持向量机的不敏感损失函数, 将支持向量机的二次寻优变为一次寻优, 加快求解速度^[8].

设训练集为 (x_i, y_i) , $i = 1, 2, \dots, n$, n 表示训练样本数, $x_i \in R^m$ 为样本输入, $y_i \in R$ 为样本输出, LSSVM 在高维特征空间的线性函数为

$$f(x) = w^T \varphi(x) + b \quad (4)$$

式中, w 为权值向量, b 为偏置量.

根据结构风险最小化原则, 综合考虑拟合误差和函数复杂度, 式(4) 变为如下的约束优化问题:

$$\min_{w, b, e} J(w, e) = \frac{1}{2} W^T W + \frac{\gamma}{2} \sum_{i=1}^n e_i^2 \quad (5)$$

$$\text{s. t. } y_i = w^T \varphi(x) + b + e_i, i = 1, 2, \dots, l$$

式中, γ 为惩罚参数; e_i 为预测误差.

通过引入 Lagrange 乘子 (Lagrange multiplier) 将式(5) 变成无约束对偶优化问题, 即

$$L(w, b, e, \alpha) =$$

$$J(w, e) - \sum_{i=1}^l \alpha_i (w^T \varphi(x_i) + b + e_i - y_i) \quad (6)$$

式中, α_i 为拉格朗日乘子.

根据 KKT 条件, 可到

$$\begin{bmatrix} 0 & e1^T \\ e1 & Q + \gamma^{-1} I \end{bmatrix} \begin{bmatrix} b \\ a \end{bmatrix} = \begin{bmatrix} 0 \\ y \end{bmatrix} \quad (7)$$

其中

$$\begin{cases} y = (y_1, y_2, \dots, y_l)^T \\ e1 = (1, 1, \dots, 1)^T \\ a = (a_1, a_2, \dots, a_l)^T \end{cases} \quad (8)$$

式中, I 表示单位矩阵, $Q_{ij} = (\varphi(x_i) \varphi(x_j))$.

对非线性分类问题, 通过引入核函数转换成线性分类问题, 相对于其他核函数, 径向基核函数参数少, 且性能更好^[9]. 本文选择径向基核函数作为 SVM 的核函数, 其定义为

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

式中, σ 为径向基核函数的宽度参数。

最后, LSSVM 分类决策函数为

$$f(x) = \text{sgn}\left(\sum_{i=1}^n \alpha_i \exp\left(-\frac{\|x_i - x_j\|^2}{2\sigma^2}\right) + b\right) \quad (10)$$

对 LSSVM 分类原理分析可知, LSSVM 学习性能与参数 γ 和 σ 相关, γ 值太大会导致分类准确率在训练过程中很高, 而在测试过程中很低, 反之则分类准确率不可靠, 模型无效; 核参数 σ 对分类结果更重要, 因为它的值影响在特征空间的分割结果: 值过大会导致过拟合, 反之不成比例的过小值则会导致欠拟合。

3 MCS 算法优化 LSSVM 参数

3.1 参数范围预估

对全部训练样本, 限制 σ 的范围:

$$k_1 d_{\min} = \sigma_l < \sigma < \sigma_u = k_2 d_{\max} \quad (11)$$

式中, $d_{\min}$ 和 $d_{\max}$ 分别表示训练集的平均最近和最远距离。

(1) 下界 σ_l . 小于 0.003 (小概率事件) 的支持向量难以覆盖其相邻样本, 即当前样本对其相邻样本影响很微小, 因此核函数的半径不宜小于此值, 否则易造成过拟合, 令 $k(x_i, x_j) = 0.003$, $\sigma_l = 0.2935 d_{\min}$.

(2) 上界 σ_u . 支持向量对距离较近样本比较远的影响大, 因此 σ 越大, 其覆盖范围相应大, $\sigma_u = 12.29 d_{\max}$.

据以上分析, 近似取 $k_1 = 0.3$, $k_2 = 13$, σ 参数的设置为 $[0.001, 1.000]$.

参数 γ 的范围不易确定, 根据经验设为 $\gamma \in [\gamma_l, \gamma_u] = [0.1, 100]$.

3.2 MCS 算法优化 LSSVM 参数

(1) 收集入侵样本, 并对入侵检测样本进行预处理, 得到 LSSVM 的学习样本。

(2) 设置 γ 、 σ 范围及 MCS 相关参数。

(3) 随机产生 n 个鸟巢位置: $p_i^{(0)} = [x_1^{(0)}, x_2^{(0)}, \dots, x_n^{(0)}]^T$, 每一个鸟巢位置对应一组参数 (γ, σ) , 计算每组参数的交叉验证误差, 并找到当前最优鸟巢 $x_b^{(0)}$ 及最小误差 $F_{\min}$ 。

(4) 保留上代最优鸟巢位置 $x_b^{(0)}$, 并根据式(2)和式(3)计算利维飞行步长, 根据利维飞行更新其它鸟巢位置, 得到一组新的鸟巢位置, 并计算它们的交叉验证误差。

(5) 根据交叉验证误差与上一代的鸟巢位置 p_{i-1} 进行比较, 用较好鸟巢位置替代较差的鸟巢位置, 得

到一组较优鸟巢位置 $k_i = [x_1^{(i)}, x_2^{(i)}, \dots, x_n^{(i)}]^T$ 。

(6) 用 r 与 P_a 比较, 保留 k_i 中被发现概率较小的鸟巢, 并更新被发现概率较大的鸟巢, 计算新鸟巢位置对应的交叉验证误差, 并与 k_i 中每个鸟巢位置交叉验证误差比较, 用较好的鸟巢位置替换较差的鸟巢位置, 得到一组新的较优鸟巢位置 p_i 。

(7) 找出步骤(7)最优鸟巢位置 $x_b^{(i)}$, 并判断其交叉验证的 $F_{\min}$ 是否满足要求, 如果满足, 则停止搜索, 并输出全局最小误差 $F_{\min}$ 以及对应最优鸟巢 $x_b^{(i)}$, 反之, 则返回步骤(4)继续寻优。

(8) 根据最优鸟巢位置 $x_b^{(i)}$ 相对应的 C 、 σ 值为 LSSVM 的最优参数。

3.3 网络入侵检测器

LSSVM 是二元分类器, 对于多种入侵类型的网络入侵检测, 必须构造多元分类器才能进行入侵检测, 本文采用一种“一对多”的方式构建多分类器, 将一种入侵类型下的样本与其余类型样本组合, 共构建 4 个 MCS-LSSVM 分类器识别 4 种常见入侵, 最终结果由分类距离最大的向量机决定, 定义有人入侵为 +1, 正常为 -1, 如图 1 所示。

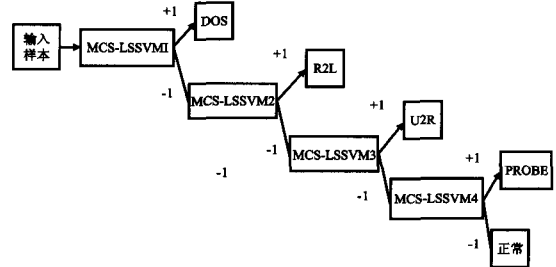


图1 多分类的网络入侵检测器

4 仿真实验

4.1 数据来源

为了检验 MCS-LSSVM 性能, 采用 KDD CUP99 数据库中的数据进行仿真实验, 该数据库包括拒绝服务攻击(DOS); 远程用户未授权访问攻击(U2R); 未授权使用本地超级权限访问攻击(U2L); 扫描攻击(Probe)等 4 种入侵方式。由于原始数据集过于庞大, 本实验选择小部分的数据集作为实验数据, 随机选择 4 000 个组合训练集和 1 000 个组成验证集。为了使 MCS-LSSVM 检测结果具有可比性, 采用遗传算法优化 LSSVM(GA-LSSVM)、粒子群优化算法优化 LSSVM(PSO-LSSVM)进行对比实验, 在相同的实验条件下, 采用检测率(%)和训练时间(秒, s)作为评价指标。

4.2 数据归一化

在 KDD CUP 99 数据集中,样本特征范围各异,而 LSSVM 是根据样本与超平面距离来确定样本的归属,因此少量特征对 SVM 性能影响大,其他特征影响较小,因此需要进行归一化处理来消除这种不利影响,具体为

$$x_i = \frac{x_i - x}{x_{\text{std}}}, i = 1, 2, \dots, n \quad (12)$$

式中, n 表示训练样本数, x_{std} 表示标准差.

4.3 结果与分析

将训练集输入到 LSSVM 中进行学习,并分别采用 GA、PSO、MCS 对 LSSVM 参数进行优化,然后采用最优参数,建立基于 GA-LSSVM、PSO-LSSVM 和 MCS-LSSVM 的网络入侵检测模型,并对验证集进行检测,检测结果见表 1.

表 1 各模型的检测性能对比

入侵类型	GA-LSSVM		PSO-LSSVM		MCS-LSSVM	
	检测率/%	训练时间/s	检测率/%	训练时间/s	检测率/%	训练时间/s
PROBE	89.81	3.41	90.04	3.23	91.74	2.04
DOS	87.45	3.68	87.47	3.02	92.97	1.86
U2R	89.94	3.15	90.26	2.92	94.24	2.22
U2L	90.92	3.38	91.24	2.66	92.80	2.31

从表 1 可以得知,采用 MCS-LSSVM 对入侵进行检测,其检测率明显要高于对比模型 GA-LSSVM、PSO-LSSVM,而且检测效率更高,训练时间明显缩短.

为了说明本文对 CS 算法改进的效果,采用基本 CS 算法优化 LSSVM 进行对比仿真实验,得到结果见表 2.

表 2 CS 算法改进前后的性能对比

入侵类型	CS-LSSVM			MCS-LSSVM		
	检测率/%	训练时间/s	支持向量数	训练时间/s	检测率/%	支持向量数
PROBE	91.15	3.07	200	2.04	91.74	189
DOS	92.74	2.81	146	1.86	92.97	138
U2R	93.87	2.57	185	2.12	94.24	181
U2L	92.22	3.14	29	2.81	92.80	28

从表 2 可知, MCS-LSSVM 检测率要高于 CS-LSSVM,而且训练时间明显减少,大幅度加快了入侵检测速度,这主要是 MCS-LSSVM 找到了更优的 LSSVM 参数,减少了支持向量点数量和样本存储容量,加快了收敛速度,更加符合现代大规模网络入侵实性、在线性检测要求.

5 结束语

针对在网络入侵检测中的 LSSVM 参数优化难题,提出一种 MCS 算法优化 LSSVM 参数的网络入侵检测模型,仿真结果表明, MCS 算法较好的解决了 LSSVM 参数优化难题,不仅可以提高网络入侵检测率,而且网络入侵检测效率得到提高,该算法是可行和有效的.

参考文献:

- [1] Heuns V, Ray H. Intrusion detection techniques and approaches[J]. Computer Communications, 2002, 25(15):1356-1584.
- [2] Alomo E J, Domingues E, Luque R M, et al. An intrusion detection system based on hierarchical self-organization[J]. Journal of Information Assurance and Security, 2009, 4(3):209-216.
- [3] Jiang D B, Yang Y H, Xia M. Research on intrusion detection based on an improved som neural network [C]//Proceedings of Fifth International Conference on Information Assurance and Security. China: Xi'an, 2009:400-403.
- [4] 李旭芳,王士同. 基于 QPSO 训练支持向量机的网络入侵检测[J]. 计算机工程与设计, 2008, 29(1):34-36.
- [5] Yang X S, Deb S. Engineering optimization by cuckoo search[J]. International Journal of Mathematical Modeling and Numerical Optimization, 2010, 1(4):330-343.
- [6] Avlyukevich I. Cooling down lévy flights[J]. Journal of Physics A: Mathematical and Theoretical, 2007(40):225-232.
- [7] Alton S, Hassan O, Morgan K, et al. Modified cuckoo search: a new gradient free optimization algorithm[J]. Chaos, Solitons & Fractals, 2011, 44(9):710-718.
- [8] 杨奎河,单甘霖,赵玲玲. 基于最小二乘支持向量机的汽轮机故障诊断[J]. 控制与决策, 2007, 22(7):778-782.
- [9] 刘智国,张雅明,林立忠. 基于粒子群 LSSVM 的网络入侵[J]. 计算机仿真, 2011, 27(11):136-139.

作者简介:

- 陈健男, (1979-), 硕士, 讲师. 研究方向为计算机应用.
- 陈雪刚男, (1976-), 硕士, 副教授. 研究方向为信息安全、数据挖掘.
- 张家录男, (1964-), 教授. 研究方向为智能信息处理、数据挖掘.
- 程杰仁男, (1974-), 博士, 副教授. 研究方向为网络安全、数据挖掘.