改进粒子群算法和支持向量机的网络入侵检测①

陶 琳, 郭春璐

(河南工业职业技术学院 电子信息工程系, 南阳 473000)

摘 要: 网络入侵检测一直是网络安全领域中的研究热点, 针对分类器参数优化难题, 为了提高网络入侵检测准确性, 提出一种改进粒子群算法和支持向量机相融合的网络入侵检测模型(IPSO-SVM). 首先将网络入侵检测率作为目标函数, 支持向量机参数作为约束条件建立数学模型, 然后采用改进粒子群算法找到支持向量机参数, 最后采用支持向量机作为分类器建立入侵检测模型, 并在 Matlab 2012 平台上采用 KDD 999 数据进行验证性实验. 结果表明, IPSO-SVM 解决了分类器参数优化难题, 获得更优的网络入侵分类器, 提高网络入侵检测率, 虚警率和漏报率大幅度下降.

关键词: 网络入侵; 特征子集; 入侵检测分类器; 支持向量机

Network Intrusion Detection Based on Improved Particle Swarm Optimization Algorithm and Support Vector Machine

TAO Lin, GUO Chun-Lu

(Department of Electronics and Information Engineering, Henan Polytechnic Institute, Nanyang 473000, China)

Abstract: Network intrusion detection is a hot research topic in network security, in order to improve the accuracy of network intrusion detection, a network intrusion detection model (IPSO-SVM) is proposed based on improved particle swarm optimization algorithm and support vector machine to solve the problem of classifier's parameters optimization. Firstly, network intrusion detection rate is taken as the objective function, and support vector machine parameters are used as the constraint conditions to establish mathematical model, and secondly improved particle swarm optimization algorithm is used to find the optimal parameters, finally, support vector machine is used as classifier to build intrusion detection model, and KDD 1999 data is used to validate the performance in Matlab 2012. The results show that IPSO-SVM has solved the optimization problem of the classifier's parameters and improved detection rate, reduced false alarm rate, false negative rate of the network intrusion.

Key words: intrusion detection; feature subset; intrusion detection classifier; support vector machine

在网络通信过程中, 需求对网络行为进行快速、有效识别, 从而保证网络的信息传递的安全性, 其中网络入侵行为检测便是其中一种识别技术[1]. 近阶段网络入侵手段呈现多样化、入侵频率越来越高, 并且给人们带来不同程度的损失, 因此提高网络入侵检测的准确性、保护网络安全具有现实意义[2,3].

网络入侵是指试图破坏计算机和网络系统资源完整性、机密性或可用行的行为, 入侵检测一些键点的信息进行搜集和分析, 从中发现网络中是否有入侵现象[4]. 对于网络入侵检测问题, 学者和专家花费了大量的时间和精力进行了广泛、深入的研究, 取得了许多研究成果[5]. 入侵检测实际就是一种多分类问题, 将网络行为划分为正常、入侵行为, 包括特征选择和入侵行为分类器两个内容, 本文针对网络入侵检测的分类器构建研究. 当前网络入侵检测的分类器主要基于神经网络和支持量机等机器学习算法进行构建[5-7], 然而神经网络基于"经验风险最小原则", 非线性建模能力强, 但训练样本规模要求大, 若不能满足该条件,

① 收稿时间:2015-10-12;收到修改稿时间:2016-01-15 [doi: 10.15888/j.cnki.csa.005304]

网络入侵检测的效果差[8]; 支持向量机专门针对小样本分类问题, 泛化能力好, 因此在网络入侵检测中应用范围更广[9]. 当支持向量机应用于网络入侵检测中时, 其检测结果与参数的选择密切相关, 当前主要采用遗传算法、粒子群算法解决支持向量机的参数选择和优化问题[10-12], 粒子群算法的搜索能力更强, 而且需要设置的参数更少, 但后其种群多样性退化严重、易陷入局部最优的缺陷[13].

为了提高网络入侵检测的准确性, 针对分类器参数优化难题, 提出一种改进粒子群算法(improved particle swarm optimization, IPSO)和支持向量机(support vector machine, SVM)的网络入侵检测模型(IPSO-SVM), 并采用 KDD 1999 数据对其性能进行分析. 结果表明, IPSO-SVM 解决了分类器参数优化难题, 获得更优的网络入侵分类器, 提高网络入侵检测率, 虚警率和漏报率大幅度下降, 可以满足网络安全的实际应用要求.

1　改进粒子群算法

在 PSO 算法中, “粒子”代表待求解问题的可行在解, 其均有位置和速度, 以一定速度在搜索空间中飞行, 设粒子的位置和位置分别为 $X=(X_1,X_2,\ldots,X_n)$和 $V=(V_1,V_2,\ldots,V_n)$, 粒粒子和粒子群最优位置分别为 $P_{i\text{best}}$和 P_{gbest}, 子的飞行过程实质就是问题的求解过程, 粒子位置和速度的更新公式如下[14]:

$$\begin{cases} v_i^{t+1}=\omega v_i^t+c_1r_1(p_{ibest}^t-x_i^t)+c_2r_2(p_{gbest}^t-x_i^t) \\ x_i^{t+1}=x_i^{t+1}+v_i^{t+1} \end{cases} \tag{1}$$

式中, t 为迭代代数; c_1、c_2 为加速系数; r_1、r_2 为(0,1)内的随机数; ω 是惯性权重.

为了提高 PSO 算法的收敛效率, 通常情况下 ω 随迭代次数增加不断减小, 具体如下

$$\omega=\omega_{\min}+\frac{(iter_{\max}-iter_{\text{i}})\times(\omega_{\max}-\omega_{\min})}{iter_{\max}} \tag{2}$$

其中, $iter_{\max}$ 和 $iter_{\text{i}}$ 分别表示最大和当前迭代次数.

在 PSO 算法工作过程中, 其存在一些缺陷, 主要为:

(1) 工作后期震荡非常严重, 收敛速度慢, 搜索效率低;

(2) 若整体收敛速度快, 种群中的个体趋同性强, 种群的多样性差, 易陷入局部最优解.

为了解决 PSO 算法存在的不足, 本文对其进行相应改进, 具体如下:

(1) 针对缺陷(1), 通过引入动量项减弱后期震荡, 加快算法的收敛速度;

(2) 针对缺陷(2), 当粒子群进行更新时, 不仅追随 $P_{i\text{best}}$和 P_{gbest}, 同时在追随群中选择某个粒子的个体极值 P_{i}, 缓和种群的趋同性.

在式(1)引入了新的学习因子 c_3 和随机量 r_3, 从而增加追随随机粒子的 P_{i}后, 得到:

$$\begin{cases} v_i^{t+1}=\omega v_i^t+c_1r_1(p_{ibest}^t-x_i^t)+c_2r_2(p_{gbest}^t-x_i^t)+ \\ \qquad c_3r_3(p_i^t-x_i^t) \\ x_i^{t+1}=x_i^{t+1}+v_i^{t+1} \end{cases} \tag{3}$$

引入动量项:

$$\begin{aligned} \Delta v_i^t=&c_1r_1(p_{ibest}^t-x_i^t)+c_2r_2(p_{gbest}^t-x_i^t) \\ &+c_3r_3(p_i^t-x_i^t) \end{aligned} \tag{4}$$

这样粒子速度和位置更新方式变:

$$\begin{cases} v_i^{t+1}=\omega v_i^t+\Delta v_i^t+\alpha\Delta v_i^{t-1} \\ x_i^{t+1}=x_i^{t+1}+v_i^{t+1} \end{cases} \tag{5}$$

式中, α 是动量因子; $\omega v_i^t+\Delta v_i^t$ 为修正量; $\alpha\Delta v_i^{(t-1)}$ 为动量项.

当 Δv_i^t 与前次同号时, 适当 $v_i^{(t+1)}$ 的速度, 提高搜索效率; 当 Δv_i^t 与前次符号相反, 减少 $v_i^{(t+1)}$ 的速度, 稳定搜索过程; 在粒子更新过程中, 采用 $P_{i\text{best}}$、P_{gbest} 和 P_i 向下一代传递信息, 增大了粒子所获得的信息量, 可以避免了算过分早熟.

选择 2 个标准函数测试本文对 PSO 算法改进的有效性进行测试, 2 个标准函数具体如下:

$$f_1(x)=\sum_{i=1}^{n}(x_i^2-10\cos(2\pi x_i)+10) \tag{6}$$

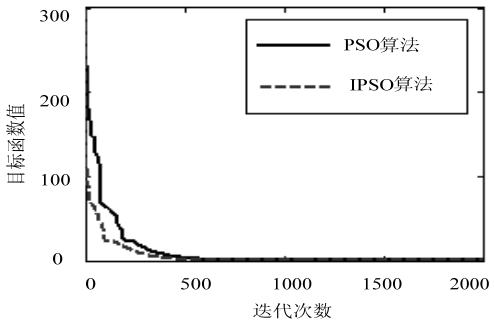
$$f_2(x)=\frac{1}{4000}\sum_{i=1}^{n}x_i^2-\prod_{i=1}^{n}\cos(\frac{x_i}{\sqrt{i}})+1 \tag{7}$$

PSO 算法和 IPSO 算法的测试结果见表 1. 从表 1 可以知道, 对于 2 个测试函数, IPSO 算法的结果均明显优于 PSO 算法, 具有更好的收敛性, 能够有效地避免早熟情况.

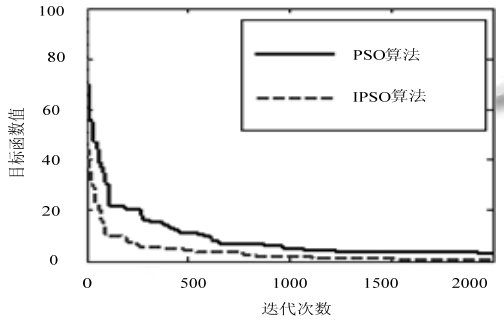
2　支持向量机

支持向量机(SVM)是一种基于统计学习理论的模式识别方法, 相对于其它机器学习算法, 支持向量机

较好的避免了“维数灾”难题的出现, 泛化能力更优, 其结构如图 2 所示[15].



(a) f_1 函数收敛曲线



(b) f_2 函数收敛曲线

图 1 PSO 算法的改进有效性测试

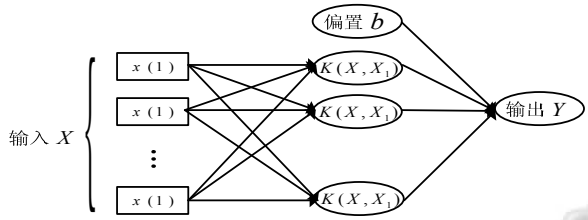


图 2 支持向量机的结构

支持向量机(SVM)通过找到一个最优分类超平面将所有训练样本划分两类, 即

$$y_i\{[\psi(x_i),\omega]+b\}\geq 1, i=1,2,...,n \tag{8}$$

式中, n 表示训练样本的数量.

对于一个分类超平面, 参数(ω,b)不唯一确定, 因此一定有一对(ω,b)保证式(8)成立, 设$\psi(x_i),y_i]$与分类超平面最小距离为$1/\|\omega\|$, 允许存在一些误分类的点, 这样式(8)变为

$$y_i\{[\psi(x_i),\omega]+b\}\geq 1-\zeta_i, i=1,2,...,n \tag{9}$$

式中, $\zeta_i(i=1,2,...,n)$ 为负松弛变量, ζ_i=0 时, 表示完全线性可分.

对于一个线性不可分问题, 那么就需要将其转为一个优化问题, 再找到最优分类超平面, 通过引入惩罚因子 C>0, 则有

$$\min\psi(\omega)=\frac{1}{2}\|\omega\|^2+C\sum_{i=1}^{n}\zeta_i$$

$$s.t. \tag{10}$$

$$\begin{cases} y_i\{[\psi(x_i),\omega]+b\}\geq 1-\zeta_i \\ i=1,2,...,n \end{cases}$$

引入 Lagrange 算子 α_i 将式(10)转化:

$$maxW(\alpha)=\sum_{i=1}^{n}\alpha_i-\frac{1}{2}\sum_{i,j=1}^{n}\alpha_i\alpha_j y_i y_j\psi(x_i)\psi(y_j)$$

$$s.t. \tag{11}$$

$$\begin{cases} 0\leq\alpha_i\leq C \\ \sum_{i=1}^{n}\alpha_i y_i=0 \end{cases}$$

式中, $\omega=\sum_{i=1}^{l}\alpha_i\psi(x_i)$.

支持向量机的分类判别函数为

$$f(x)=\omega\psi(x)+b=\sum_{i\in SV}\alpha_i\psi(x_i)\psi(x)+b \tag{12}$$

特征空间的内积(核函数)为

$$K(x,y)=\sum_i\psi_i(x)\psi_j(y) \tag{13}$$

高斯核定义为

$$K(x,y)=\exp(-\frac{\|x-y\|^2}{2\sigma^2}) \tag{14}$$

式中, σ为高斯分布宽度.

支持向量机的分类判别函数为

$$f(x)=\omega\psi(x)+b=\sum_{i\in SV}\alpha_i K(x_i,x)+b \tag{15}$$

3 PSO-SVM的网络入侵模型

3.1 SVM 参数优化的数学模型

采用一个分类问题, 选择不同的 C、σ 分析它们对 SVM 分类性能影响, 具体见表 1 和表 2. 由表 1 和 2 可以发现, 当 C、σ 的取值不同时, SVM 的分类性能差别很大, 这说明 C 和 σ 值的优劣直接决定了 SVM 的分类性能, 因此要建立最优的 SVM 网络入侵检测模型, 要解决 SVM 参数的优化问题.

表 1　C 对分类 SVM 的性能作用

C	分类精度(%)
20.42	53.62
82.66	61.25
185.96	78.96
226.57	61.20
447.99	89.17
886.90	79.59

表2　σ对SVM的分类性能作用

σ	检测率(%)
0.181	67.704
0.545	53.147
0.699	75.531
1.106	65.256
1.471	71.601
9.570	76.06

SVM 参数优化的目标是寻找最佳的参数组合, 使得网络入侵的检测率最高, 因此SVM参数的优化的数学模型为:

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

3.2 PSO-SVM 的入侵检测步骤

(1) 收集网络状态历史数据, 并进行相应预处理.

(2) 随机产生一组初始粒子, 每一个粒子位置串包括 C、σ.

(3) 对粒子位置串进行反编码, 根据适应度值确定个体极值和群体极值.

(4) 更新粒子的速度和位置产生新粒子群.

(5) 对粒子位串进行解码, 并计算粒子适应度值.

(6) 若达到最大迭代次数, 那么就返回全局最优粒子位置, 若不满足跳转至步骤(4)循环.

(7) 将最优粒子位置解码成为 C、σ 并建立最优网络入侵检测模型.

4　仿真实验

4.1 数据源

选择 KDD CUP 99 数据集测试 IPSO-SVM 的有效性与优越性, 该数据来源于 DARPA 在 MIT 林肯实验室, 包括 4 类入侵类型: DOS, U2R, R2L, Probe 和正常行为(Normal), 每一个连接通过 41 个特征进行描述, 随机选择 1000 个样本作为实验对象, 具体如表 3 所示, 选择 Matlab 2012 平台实现仿真实验.

表 3　样本分布

网络状态	样本数集
DoS	200
Probe	200
R2L	70
U2R	30
Normal	500

4.2 对比模型以及参数选择

选择采用遗传算法优化 SVM(GA-SVM)和 PSO 优化 SVM(PSO-SVM)作为对比模型, 选择检测率、虚警率、漏报率评价网络入侵检测的结果, 它们具体如下:

$$\text{准确率}=\frac{\text{检测准确的样本数}}{\text{样本总数}}\times 100\% \tag{17}$$

$$\text{虚警率}=\frac{\text{被误报为入侵的正常样本数}}{\text{正常样本总数}}\times 100\% \tag{18}$$

$$\text{漏报率}=\frac{\text{被误报为正常的正常样本数}}{\text{入侵样本总数}}\times 100\% \tag{19}$$

采用 GA、PSO 以及 IPSO 算法对 SVM 参数进行寻优, 得到的结果如表 4 所示.

表 4　SVM 的参数选择结果

网络状态	GA		PSO		ISPO	
	C	σ	C	σ	C	σ
DoS	938.96	7.83	982.05	0.10	395.82	6.47
Probe	308.79	1.07	956.66	6.17	865.79	7.13
R2L	318.42	3.85	127.41	6.53	110.94	0.57
U2R	252.52	8.29	462.49	2.73	999.57	4.21
Normal	850.14	9.82	90.96	9.68	116.26	4.65

4.3 结果与分析

采用表 4 的参数建立相应的网络入侵检测模型, GA-SVM、PSO-SVM 和 IPSO-SVM 的检测结果统计如表5所示. 从表5的检测结果可以发现, 相对于GA-SVM以及PSO-SVM, IPSO-SVM获得更高的入侵检测检测率, 平均检测率分别比 SO-SVM, IPSO-SVM 要高 8.102%和 6.33%, 同时虚警率、漏报率均得到了不同程度的降低, 这说明 IPSO 算法获得了比 GA 以及 PSO 算法更优的 SVM 参数 C, σ, 建立了更加理想的网络入侵检测模型, 验证了 IPSO-SVM 的有效性与优越性.

表 5　入侵检测结果的对比

模型	评价指标(%)	DoS	Probe	R2L	U2R	Normal
GA-SVM	准确率	82.67	77.60	89.72	91.84	92.52
	虚警率	11.28	19.15	6.76	7.51	6.37
	漏报率	6.05	3.25	3.52	4.65	2.11

	准确率	86.53	79.70	90.93	92.75	93.30
PSO-SVM	虚警率	9.94	15.56	4.73	2.27	3.05
	漏报率	3.53	2.74	2.34	2.98	1.65
	准确率	94.72	83.47	97.69	98.98	100
IPSO-SVM	虚警率	4.81	10.86	1.57	1.32	0
	漏报率	0.47	5.67	0.74	0.3	0

GA-SVM、PSO-SVM 和 IPSO-SVM 的平均检测时间统计如图 3 所示. 从图 3 可知, 相对于 PSO-SVM 和 IPSO-SVM, IPSO-SVM 的网络入侵检测时间更少, 主要由于IPSO算法可以更快找到支持向量机参数, 训练过程中的计算复杂度下降, 提高了网络入侵检测的效率.

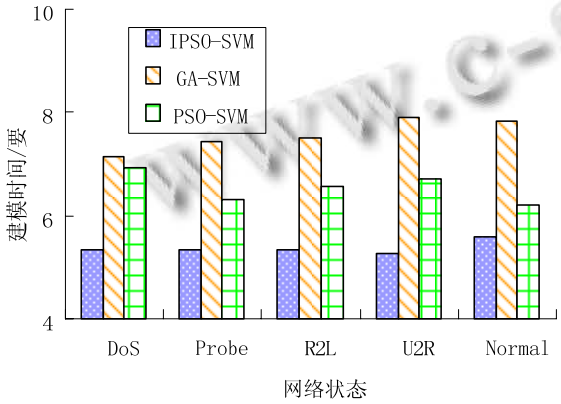


图 3 平均建模时间对比

5 结束语

提高网络入侵检测性能具有十分重要的实际应用价值, 要改善网络入侵的检测性能, , 提出一种改进粒子群算法和支持向量机的网络入侵检测模型, 采用 KDD 999 数据集对 IPSO-SVM 的性能进行了测试, 可以得到如下结论:

(1) 针对支持向量参数选择的难题, 通过改进粒子群算法对其解决, 得到合理的支持向量机, 建立更优的入侵检测分类器, 提高了网络入侵检测率.

(2) 针对 PSO 算法在支持向量参数寻优过程中收敛速度慢、易陷入局部最优的缺陷, 通过引入动量项减弱后期震荡加快收敛速度, 并对粒子群更新方式进行改进, 增加种群的多样性性, 降低了网络入侵检测的误检率和漏检详细, 而且入侵检测结果要优于其它算法, 在网络安全管理中有着广泛的应用前景.

参考文献

1 Patel R, Thakkar A, Uanatra A. A survey and comparative analysis of data mining techniques for network intrusion detection systems. International Journal of Soft Computing Engineering, 2012, 2(1): 78−85.

2 严岳松,倪桂强,缪志敏等.基于SVDD的半监督入侵检测研究.微电子学与计算机,2012,26(10):128-130

3 Dong Y, Qi B, Zhu W, et al. A new intrusion detection model based on data mining and neural network. Przeglad Elektrotechniczny, 2013, 89(1b): 88-90.

4 Denning DE. An intrusion detection model. IEEE Trans. on Software Engineering, 2010, 13(2): 222−232.

5 Pan W, Shen XT, Liu BH. Cluster analysis unsupervised learning via supervised learning with a non convex penalty. Journal of Machine Learning Research, 2013: 1865−1889.

6 朱红萍,巩青歌,雷战波.基于遗传算法的入侵检测特征选择.计算机应用研究,2012,29(4):1417−1419.

7 Wang B, Shi Y, Huang WW, et al. Misclassification minimization based on multiple criteria linear programming. Proc. of 2014 IEEE Int Conl on Data Mining Workshop, Piscataway, NJ. IEEE. 2014. 88−92

8 李超,李文法,段洣毅.用于网络入侵检测的 VFSA-C4.5 特征选择算法.高技术通讯,2011,21(12):1420−1425.

9 Venkatesan R, Uanesan R, Selvakumar AAI. A comprehensive study in data mining frameworks for intrusion detection . International Journal 01 Advanced Computer Research, 2012, 2(7): 29−34.

10 马世欢,胡彬.基于特征选取和样本选择的网络入侵检测.计算机系统应用,2015,24(9):240−243.

11 李振刚,甘泉.改进蚁群算法优化 SVM 参数的网络入侵检测模型研究.重庆邮电大学学报(自然科学版),2014,26(6):785−789.

12 王雪松,梁昔明.改进蚁群算法优化支持向量机的网络入侵检测.计算技术与自动化,2015,2:95−99.

13 张拓,王建平.基于 CQPSO-SVM 的网络入侵检测模型.计算机工程与应用,2015(2):113−116.

14 李欣然,靳雁霞.权重自适应调整的混沌量子粒子群优化算法.计算机系统应用,2012,21(8):127−130.

15 肖国荣.改进蚁群算法和支持向量机的网络入侵检测.计算机工程与应用,2014(3):75−78.