

文章编号:1001-9081(2010)06-1486-03

基于遗传进化和粒子群优化算法的入侵检测对比分析

郑洪英¹,倪霖²,侯梅菊¹,王渝¹

(1. 重庆大学 计算机学院, 重庆 400030; 2. 重庆大学 机械工程学院, 重庆 400030)

(zhenghongy@cqu.edu.cn)

摘要:针对入侵检测中的聚类最优化问题,使用遗传算法和粒子群算法的优化特性进行全局最优化并作对比分析。分析采用二进制编码,终止条件同时考虑最大迭代次数和收敛度,适应度函数的定义结合了类内距和类间距的特征。最后使用 KDD CUP1999 数据集在 Matlab 6.5 中进行了仿真。实验结果表明粒子群算法在适应度的收敛值和收敛速度上均优于遗传算法。

关键词:遗传算法;粒子群算法;入侵检测;聚类;进化;最优化

中图分类号: TP393.08 **文献标志码:** A

Comparative study on evolutionary genetic algorithm and particle swarm optimization in intrusion detection

ZHENG Hong-ying¹, NI Lin², HOU Mei-ju¹, WANG Yu¹

(1. College of Computer Science, Chongqing University, Chongqing 400030, China;

2. College of Mechanical Engineering, Chongqing University, Chongqing 400030, China)

Abstract: Concerning the clustering optimization in intrusion detection, Genetic Algorithm (GA) and Particle Swarm Optimization (PSO) were used to optimize clustering and comparative analysis was also completed. In this analysis, binary code was adopted and elimination criterion took the account of the maximum number of iteration and the quality of convergence. Fitness function which combines the characteristic of inter-cluster distance and intra-cluster distance was defined. Finally, the experiments with KDDCUP 1999 data set using Matlab 6.5 tools show that PSO is superior to GA in the value and speed of fitness function convergence.

Key words: genetic algorithm; particle swarm optimization; intrusion detection; clustering; evolution; optimization

0 引言

入侵检测技术是安全防御体系的一个重要组成部分,是把异常数据从正常数据中抽取出来,因此入侵检测问题可以转化成数据的最优分类问题,也就是找到正确分类数据的最优解。而进化优化算法是模拟自然进化过程,根据适应度大小进行个体的优胜劣汰,逐步逼近最优解。针对复杂问题,进化优化算法具有很强的搜索能力和最优化性能,所以可以使用进化优化算法来解决入侵检测中数据分类问题以及和分类有关的参数确定问题,以此来提高检测率和降低误报率。

遗传算法是一种通过模拟自然进化过程搜索最优解的方法,对一个由多个解构成的种群进行评估、遗传运算、选择,经多代繁殖,获得适应值最好的个体作为问题的最优解。文献[1-2]使用混合遗传算法优化聚类结果达到了较好效果。而粒子群算法^[3-4]在连续优化问题、组合优化问题等方面的应用也较广泛。文献[5]使用 K-PSO 算法进行入侵检测,克服了 K 均值算法的缺点,提高了检测效率。因此,本文以提高入侵检测的准确率和降低误报率为出发点,针对入侵检测中聚类分析的全局最优化问题,分析遗传算法和粒子群算法的应用模型,并进行两种算法优化特性和检测效果的对比分析。

1 遗传算法在入侵检测中的应用

遗传算法主要以两类方式应用于入侵检测中:一是直接使用遗传算法产生出分类规则^[6-7];二是用遗传算法来选择更适合的特征或对一些函数参数进行优化。

1.1 数据分析与染色体编码

由于正常数据与攻击数据间存在着本质的区别,因此可以根据相似度将正常数据与攻击数据分离。通过聚类确定两个聚类中心,将训练集通过算法分成正常数据和攻击数据两类。通过数据分析,可以使用二进制对训练集进行编码,1 代表正常数据,0 代表攻击数据,染色体的长度 N 即为训练集的规模。

1.2 选择和交叉方式

由于样本经过标准化后,数据间的差别变得很小,计算出来的适应值相差也很小,因此可以采用截断选择。选择最好的前 T 个个体,让每一个个体都有 $1/T$ 的选择概率,即平均每个个体得到 N/T 个繁殖机会(N 为种群规模)。截断选择确保了优秀个体能确定地遗传到下一代,劣质个体将失去繁殖的机会,体现了自然界“优胜劣汰”的原则。

鉴于该算法种群染色体长的特点,引入了基因的概念,将染色体分成一个一个的基因片段。若染色体的长度为 N ,单个基因的长度为 M ,则整个染色体被分割成 N/M 个基因。对每

收稿日期: 2009-12-23; **修回日期:** 2010-03-01。 **基金项目:** 国家 863 计划项目(2006AA04A123); 重庆市自然科学基金资助项目(2008BB2182; 2008BB0173); 重庆大学青年骨干教师创新能力培育基金资助项目(CDCX021)。

作者简介: 郑洪英(1975-),女,重庆人,讲师,博士,主要研究方向:信息安全; 倪霖(1971-),男,重庆人,副教授,主要研究方向:信息系统; 侯梅菊(1987-),女,重庆人,硕士研究生,主要研究方向:信息安全; 王渝(1982-),男,重庆人,硕士研究生,主要研究方向:信息安全。

个基因进行单点交叉,整体上看就相当于对整条染色体进行了多点交叉,确保了交叉的范围足够大,更好地产生出新的种群。

1.3 终止条件

除了将最大迭代次数作为停止准则外,还可以根据种群的收敛程度,即种群中适应值的一致性来判断是否算法停止。在算法过程中保留历史最好的个体的适应值,使用式(1)判断种群是否收敛。式(1)中, F 为最后 M 个历史最好个体的适应值, \bar{F} 为 F 的平均值, ε 是一个足够小的常量。通过控制 ε 的值,可以控制在最后 M 历史最优值的方差小于 ε 时认为算法达到收敛。

$$\sum |F_i - \bar{F}| < \varepsilon \quad (1)$$

2 粒子群算法在入侵检测中的应用

由于编码方式是离散的二进制编码^[8],因此,将使用二进制粒子群优化算法来优化数据集的聚类。

2.1 二进制粒子群算法

设粒子群有 N 个粒子,每个粒子相当于 D 维离散空间中的一个活动点。粒子 i 在 t 时刻的速度、位置、个体最好位置和群体最好位置分别用 $v_i(t)$ 、 $x_i(t)$ 、 $p_i(t)$ 和 $g_i(t)$ 表示,那么,粒子 i 的速度和位置的迭代公式如下:

$$\begin{aligned} v_{id}(t+1) &= \omega \cdot v_{id}(t) + c_1 r_1 (p_{id}(t) - x_{id}(t)) + \\ &\quad c_2 r_2 (g_d(t) - x_{id}(t)) \quad (2) \\ x_{id}(t+1) &= \begin{cases} 1, & r_3 < \text{Sig}(v_{id}(t+1)) \\ 0, & r_3 \geq \text{Sig}(v_{id}(t+1)) \end{cases} \quad (3) \end{aligned}$$

其中: $i = 1, \dots, N$ (N 表示群体大小,一般取20); $d = 1, \dots, D$ (D 表示粒子编码中各分量的维数,由具体问题给定); r_1, r_2, r_3 均为 $(0,1)$ 的随机数; c_1 和 c_2 为学习因子,通常取 $c_1 = c_2 = 2$; ω 为非负数,称为惯性因子,也叫惯性权重,在二进制粒子群优化算法中一般取 $\omega = 1$; $\text{Sig}(\cdot)$ 为sigmoid函数,通常取 $\text{Sig}(x) = 1/(1 + \exp(-x))$ 。

2.2 适应度函数

设样本集中一个类 $X = \{x_i | i = 1, 2, \dots, N\}$,其中 x_i 为 D 维向量,则 X 的类内聚 f_X 满足:

$$f_X = \sum_{i=1}^N \sum_{j=1}^N \|x_i, x_j\| \quad (4)$$

X 的聚类中心 X_c 满足:

$$X_c = \frac{1}{N} \sum_{i=1}^N x_i \quad (5)$$

若样本集中另一个类 $Y = \{y_i | i = 1, 2, \dots, M\}$,其中 y_i 为 D 维向量,则 X 和 Y 的类间距 $d(X, Y)$ 即为两个聚类中心

的欧氏距离。适应度函数结合使用类内距和类间距,同时考虑类内的聚合度和类间离散度,适应值 f' 定义为式(6)。类内距越小,类间距越大,故 f' 的值越小。通过选择较小的适应值进行运算将达到收敛的效果。

$$f' = \frac{f_X + f_Y}{d(X, Y)} \quad (6)$$

3 实验结果

实验中使用 KDD Cup1999 作为实验数据, KDD Cup1999^[9]中总共包括了41个特征,首先选取了8维数据作为训练集。其中第8维数据为标记位,它唯一确定了数据的属性。具体的8维数据属性是:src_bytes、dst_bytes、count、srv_count、dst_host_count、dst_host_srv_count、dst_host_same_src_port_rate、falg。为了消除特征量纲对实验结果的影响,连续型数据需要规范化。假设 $X = \{x_{ij} | i = 1, \dots, N; j = 1, \dots, D\}$ 是输入数据, N 是样本数据的数目, D 是样本数据的特征维数, μ 是均值, σ 是样本的标准差,则规范化数据 $x'_{ij} = (x_{ij} - \mu)/\sigma$ 。

3.1 算法参数设置

算法参数对收敛有很大影响,经过综合测试设定的参数如表1所示。

表1 算法参数设置

遗传算法参数	取值
种群规模 popsize	20
最大迭代次数 eratum	1000
种群收敛的最后 key 次迭代次数	50
基因长度 Galen	10
交叉率 pCross	0.9
变异率 pMutation	0.05
粒子群算法参数	取值
种群规模 popsize	20
最大迭代次数 eratum	1000
种群收敛的最后 key 次迭代次数	50
最大速度 v_{\max}	10
学习因子 $c_1 = c_2$	2
惯性权重 ω	1

3.2 实验结果

分别运行遗传算法聚类程序和粒子群算法聚类程序得到的检测结果如表2所示。评价指标主要包括分类准确率、检测率、误报率、漏报率和运行时间等,分类准确率即为算法所加标签与原数据集标签的匹配程度。图1~3分别给出了种群规模为20、10和5时遗传算法和粒子群算法的适应度函数收敛过程。

表2 两种算法的检测对比结果

算法	迭代次数	最优适应值	分类准确率	检测率	误报率	漏报率	运行时间
遗传算法	211	4184.695805	0.993333333	0.968	0.022727273	0.14	4 min 3 s
粒子群算法	211	4044.632502	1	0.968	0.022727273	0.14	3 min 58 s

3.3 结果分析

通过算法的运行结果可以看出,两种算法都具有很强的收敛能力,能在有限的次数(250以内)达到收敛。从图1可以看出当种群规模为20时,两种算法的收敛结果非常接近。但是随着这个参数值的降低,遗传算法收敛结果越来越不理

想,而粒子群算法在种群规模为5时仍然收敛,适应值较遗传算法更优,所以遗传算法对这个参数相对于粒子群算法而言更加敏感;从分类准确率来看,粒子群算法更具有优越性。最优适应值越好越接近训练集的最优适应值,分类的准确率越高,从而检测率也越高;算法的运行时间与算法的迭代次数有

直接的关系,迭代次数越多,运行时间越长,遗传算法与粒子群算法的运行时间基本持平;而在适应值曲线的表现上,粒子群算法的曲线较陡峭,遗传算法的曲线较平缓,正确地反应出了遗传算法和粒子群算法的特点。粒子群算法由于粒子运动在大体上是向着粒子本身的历史最优位置和群体最优位置前进,但小范围内粒子的运动是离散随机的,某个粒子偶然找到一个更好的位置,其他粒子便会向着这个位置靠拢。在适应值曲线上表现出来就显得比较陡峭,突变点较多。而遗传算法每一代都选取了适应值较好的父代进行遗传,逐步逼近最优适应值,由于每次交叉和变异对种群的改变不是很大,因此在适应值曲线上表现出来就显得平缓。

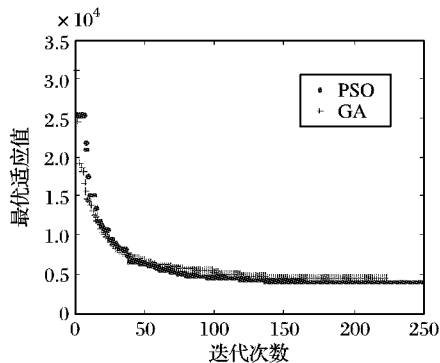


图1 种群规模 $popsiz = 20$ 时算法的收敛过程

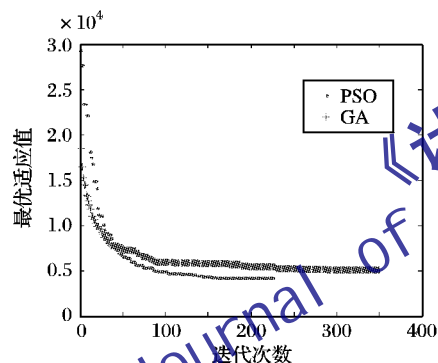


图2 种群规模 $popsiz = 10$ 时算法的收敛过程

4 结语

文中对遗传和粒子群算法在入侵检测中的优化特性进行对比分析。在遗传算法方面,引入了基因的概念,将染色体分成一个一个的基因进行交叉和变异运算,简化了算法的复杂度,且提高了交叉的范围。在粒子群优化算法方面采用了二

进制粒子群优化算法,这对于组合优化问题具有更好的适用性。在两种算法的终止条件判定方面,既使用了最大迭代次数作为终止条件,同时又使用了收敛度判断。实验结果证明:两种进化优化算法在训练集上聚出的类与原始类相差很小,进化优化算法的高效性和优异性得到了充分的体现,但是粒子群算法在适应度的收敛值和收敛速度上均优于遗传算法。

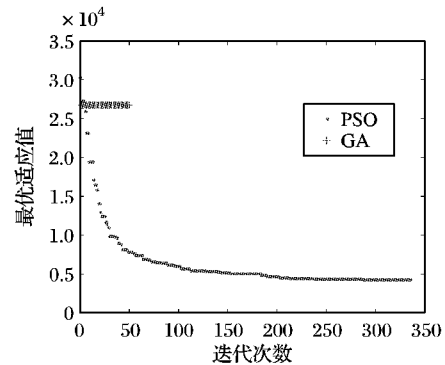


图3 种群规模 $popsiz = 5$ 时算法的收敛过程

参考文献:

- [1] 唐少先, 蔡文君. 基于无监督聚类混合遗传算法的入侵检测方法[J]. 计算机应用, 2008, 28(2): 400-411.
- [2] 徐东升, 艾晓燕, 阎世梁. 基于遗传优化与模糊规则挖掘的异常入侵检测[J]. 计算机应用, 2009, 29(8): 2227-2229.
- [3] EBERHART R, KENNEDY J. A new optimizer using particle swarm theory [C]// Proceedings of the Sixth International Symposium on Micromachine and Human Science. Washington, DC: IEEE, 1995: 39-43.
- [4] KENNEDY J, EBERHART R. Particle swarm optimization [C]// Proceedings of IEEE International Conference on Neural Networks. Piscataway, NJ: IEEE Service Center, 1995: 1942-1948.
- [5] 谷保平, 许孝元, 郭红艳. 基于粒子群优化的k均值算法在网络入侵检测中的应用[J]. 计算机应用, 2007, 27(6): 1368-1370.
- [6] 郭惠玲, 唐勇, 张冬丽. 遗传算法在入侵检测规则提取中的应用[J]. 哈尔滨工业大学学报, 2009, 41(1): 348-350.
- [7] ABADEH M S, HABIBI J, BARZEGAR Z, et al. A parallel genetic local search algorithm for intrusion detection in computer networks [J]. Engineering Applications of Artificial Intelligence, 2007, 20(8): 1058-1069.
- [8] 姜永森, 王军霞, 杨慧中. 基于二进制粒子群优化的决策系统属性离散化[J]. 控制工程, 2008, 15(4): 360-363.
- [9] KDD Cup 1999 data [DB/OL]. [2008-10-10]. <http://kdd.ics.uci.edu/data-bases/kddcup.html>.

(上接第1482页)

参考文献:

- [1] 韦勇, 连一峰. 基于日志审计与性能修正算法的网络安全态势评估模型[J]. 计算机学报, 2009, 32(4): 763-772.
- [2] 赖积保, 王慧强, 朱亮. 网络安全态势感知模型研究[J]. 计算机研究与发展, 2006, 43(增刊): 456-460.
- [3] 朱帮助, 林健. 基于ARIMA和LSSVM的非线性集成预测模型[J]. 数学的实践与认识, 2009, 39(12): 34-40.
- [4] 张翔, 胡昌振, 刘胜航, 等. 基于支持向量机的网络攻击态势预测技术研究[J]. 计算机工程, 2007, 33(11): 10-12.
- [5] 任伟, 蒋兴浩, 孙敏峰. 基于RBF神经网络的网络安全态势预测方法[J]. 计算机工程与应用, 2006, 42(3): 136-144.
- [6] SHEN LIU-QING, WANG JIN-DONG, WANG KUN, et al. The

design of intelligent security defensive software based on autonomic computing [C]// The Second International Conference on Intelligent Computation Technology and Automation. Washington, DC: IEEE Computer Society, 2009: 489-491.

- [7] 储小俊, 刘思峰. 基于新陈代谢GM-Markov模型的股价预测[J]. 山东财政学院学报, 2007, 3(1): 43-45.
- [8] 居玲华, 石培基. 基于Markov和GM(1,1)模型的土地利用结构预测[J]. 农业系统科学与综合研究, 2009, 25(2): 138-146.
- [9] PROJECT H. Know your enemy: statistics [EB/OL]. [2009-10-20]. <http://old.honeynet.org/papers/stats>.
- [10] ROESCH M, GREEN C. SNORT users manual [EB/OL]. [2009-11-02]. http://www.snort.org/assets/82/snort_manual-2_8_5_1.pdf.