DOI:10.13954/j.cnki.hdu.2008.02.023

一种改进免疫遗传算法及在异常检测中的应用

刘 董,郑 宁,徐海涛
(杭州电子科技大学计算机学院,浙江 杭州 310018)

摘要:该文针对免疫遗传算法的不足,在分析其特性的基础上,引入了隔离小生境技术,改进交叉算子和变异算子,提出一种改进算法。在基于模糊关联规则挖掘的异常检测中采用本算法优化后的隶属函数,能够扩大正常关联规则集之间的相似度,缩小正常与异常关联规则集之间的相似度,提高异常检测的性能。通过以网络流量为数据的异常检测实验仿真对算法进行了验证。实验结果说明了该算法的可行性和有效性。

关键词:异常检测;隶属函数;免疫遗传算法;隔离小生境
中图分类号:TN401 文献标识码:A 文章编号:1001-9146(2008)02-0057-04

0 引 言

如何优化隶属函数是模糊关联规则挖掘方法在网络异常检测中应用的一个难点。这是因为隶属函数是定量分析模糊属性的基础,而它的确定往往依赖于专家领域知识。目前有些学者提出用遗传算法(Genetic Algorithm, GA)对隶属函数进行优化[1-3]。但传统的遗传算法本身存在着许多不足,例如易出现未成熟收敛等[4]。为此,本文提出一种改进免疫遗传算法(Improved Immune Genetic Algorithm, IIGA),并将其应用在网络异常检测之中。

1 基于模糊关联规则挖掘的异常检测

模糊关联规则挖掘算法是异常检测中常用的一种方法。其应用的具体方法是:建立正常状态下的关联规则集 S_1;挖掘某暂态下的关联规则集 S_2;计算两规则集之间的相似度 Similarity(S_1, S_2);根据相似度来判断系统是否处于异常状态。

隶属函数的选取是模糊关联规则挖掘算法应用的一个难点。本文选取3个标准隶属函数S,PI,Z,将模糊变量分为High,Medium,Low3个模糊集合。这3个函数的定义为:

$$S(x,a,c)=\begin{cases}0 & x\leqslant a\\ 2\left(\frac{x-a}{c-a}\right)^2 & a<x\leqslant\frac{a+c}{2}\\ 1-2\left(\frac{x-a}{c-a}\right)^2 & \frac{a+c}{2}<x\leqslant c\\ 1 & c<x\end{cases} \tag{1}$$

收稿日期:2007-09-14
基金项目:浙江省自然科学基金(Y106176),浙江省科技厅科技计划项目(C33058)
作者简介:刘 董(1982-),女,浙江定海人,在读研究生,网络安全.

$$Z(x, a', c') = 1 - S(x, a', c') \tag{2}$$

$$PI(x, d, b) = \begin{cases} S(x, b-d, b) & x \leqslant b \\ Z(x, b-b+d) & b < c \end{cases} \tag{3}$$

2 IIGA 算法

免疫遗传算法是将免疫系统特性与遗传算法结合起来的一种算法。把免疫系统的识别多样性特点应用于简单遗传算法之中,可以在很大程度上克服遗传算法中常见的早熟现象[5]。但免疫遗传算法仍具有遗传算法的一些不足,为了克服上述不足,本文提出一种改进算法。

2.1 隔离小生境技术的引入

生物在进化过程中,之所以具有形成千变万化的物种的能力,主要是因为生物种群具有分化为小种群的能力。在种群分化过程中,隔离起着非常重要的作用。为了保持解的多样性,本文算法借鉴隔离遗传思想,将原始种群划分为多个子种群分别进行进化。同时对多个子种群也采取适者生存的原则,使子种群的规模同种群内个体的平均适应度相适应。子种群的平均适应度越高则该子种群的规模就越大,反之就越小。子种群 k 第 t+1 代的规模 $n_k(t+1)$为:

$$n_k(t+1) = M \times f_k(t) / \sum_{i=1}^{N} f_i(t) \tag{4}$$

式中,M 为种群规模,N 为子种群个数,$f_k(t)$为第 t 代 k 子种群的平均适应度。

2.2 改进的交叉算子和变异算子

传统的交叉算子,对于要进行交叉的两个个体的选择是随机的,这样存在两个近似个体进行交叉的现象,从而使新个体变异程度不大。本文提出一种基于范式距离的交叉个体选择方法。对于一个未经过交叉的个体,依次计算它与其他未交叉个体的范式距离 $\sqrt{(d_{i1}-d_{j1})^2+(d_{i2}-d_{j2})^2+\cdots+(d_{ik}-d_{jk})^2}$,挑选与其范式距离值最大的个体作为其交叉对象。

对于变异算子,传统的变异算子将变异率维持在一个固定的水平。这就使得在进化前期,当交叉作用显著时,变异效果不明显;而当进化后期,固定的较小的变异率使得依靠现有群体找到最优解的可能性变得较小。本文将变异率设计为:

$$P_m(g) = P_{min} + N_G a \tag{5}$$

式中,g 表示当前代数,P_{min}表示最小变异率,a 为变异率提高系数,N_G 为自上次进化以来至当前代为止连续未进化代数。未进化时间越长,则 N_G 越大,从而变异率扩大。

2.3 IIGA 的一般步骤

IIGA 的整个优化搜索过程由以下几个步骤组成:

(1)抗原的识别。选取待求解问题的函数作为抗原;

(2)随机产生初始种群并编码。将待优化参数的组合作为抗体。随机产生的 M 个抗体共同构成初始种群,并根据实际需要选取一种编码方式,对初始种群进行编码;

(3)初始种群的隔离。将种群中的 M 个个体,平均分成 N 个子种群;

(4)计算个体适应度。计算个体适应度并将各子种群中最佳个体存入免疫记忆矩阵;

(5)确定子种群规模。根据前述方法确定子种群的规模;

(6)抗体的促进与抑制。计算当前各子种群中适应度值相近的个体浓度,浓度高则减小该个体的选择概率即抑制;反之,则增加该个体的选择概率即促进;

(7)抗体产生。用改进的交叉算子和变异算子进行与标准遗传算法相似的操作;

(8)算法终止判断。采用限定迭代次数的终止判断,若满足则结束,否则重复步骤(4)—(7)。

3　IIGA 在异常检测中的应用

用 IIGA 对本文所选隶属函数进行优化,对编码方案和适应度函数的设计加以说明。

(1)编码表示

采用二进制编码来表示遗传基因的染色体。根据式 1~3,定义模糊集合的每个标准函数需要确定 2 个参数,则定义模糊变量的隶属函数共需 6 个参数。一个基因代表一个模糊变量。

(2)确定适应度函数

针对本文所研究的异常检测技术,适应度函数是建立在模糊关联规则相似度基础之上的。隶属函数优化的目标是尽量使得正常关联规则集之间的相似度最大化,正常与异常关联规则集之间的相似度最小化。本文将分别挖掘正常状态参考关联规则集 S_r,正常状态关联规则集 S_n 和异常状态关联规则集 S_a。因此将适应度函数定义为:

$$F=S_m/(1+S_{ra}) \tag{6}$$

式中,S_m=Similarity(S_r, S_n), S_m=Similarity(S_r, S_a)。

4　实验仿真

本文实验数据来自 http://ivpr.cs.uml.edu/shootout/netdoc2.htm,从该站点下载两组 tcpdump 形式的网络流量数据 base 和 net1。其中,base 代表正常状态下的网络流量数据,net1 包含了模拟攻击的异常网络流量数据。把 base 数据分成两组,分别用 reference 和 normal 标识。reference 用于训练生成 S_r,而 normal 和 net1 用来挖掘生成 S_n 与 S_a。对流量数据进行预处理,得到 3 组数值属性 TS、TF 和 TR。它们分别表示过去两秒内,在包头中包含 SYN、FIN 和 RST 标记的 TCP 包的数量。这 3 个属性也看作是模糊变量。

按照本文所述的方法随机形成 40 个个体组成初始群体,平均分成两组形成初始的隔离子种群。取最小支持度为 0.1,最小可信度为 0.6,按照模糊关联规则挖掘算法分别挖掘 S_r、S_n 和 S_a。计算相似度 S_m 和 S_{ra} 及个体适应度。取 $P_m(t)=0.04+0.01N_G$,交叉概率 $P_c=0.5$,对初始子种群分别进行免疫遗传操作,得到下一代个体。取迭代数为 50,重复操作,直到循环结束,得到最优个体。取相同的参数,用简单遗传算法进行与上述同样的优化实验。两组实验中各代最优个体所对应的适应度值如图 1 所示。

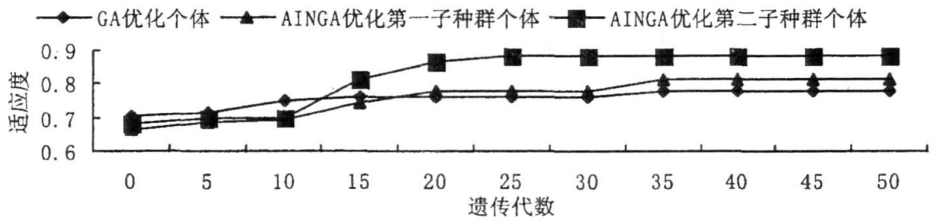

图 1　各代最优个体适应度

图 1 实验数据表明,IIGA 比 GA 能找到更佳的隶属函数,具有更好的全局寻优能力。

在种群规模为 20、40、100 个个体,相应的遗传代数分别为 30 代、40 代和 100 代的情况下,重复上述两组实验。用优化得到的最优个体进行模糊关联规则挖掘,并计算 S_m 和 S_{ra},实验结果如表 1 所示。表

1中数据表明，与采用GA优化的结果相比，在模糊关联规则挖掘中采用IIGA优化的隶属函数后，正常关联规则集之间的相似度(S_m)提高，正常与异常关联规则集之间的相似度(S_{ra})降低；最大限度地把系统正常状态与异常状态区分开来。

表1 IIGA和GA优化后各规则集的相似度

比较项	规则集相似度	20个体遗传30代	40个体遗传50代	100个体遗传100代
IIGA优化后	S_{rn}	0.943 3	0.957 7	0.986 9
	S_{in}	0.126 1	0.080 9	0
GA优化后	S_{rn}	0.789 9	0.862 1	0.875 1
	S_{in}	0.144 3	0.106 7	0.053 6

5 结 论

本文在分析免疫遗传算法特性和局限性的基础上，提出了一种改进算法，即IIGA算法。将IIGA算法应用到基于模糊关联规则挖掘的异常检测中，并以网络流量为数据进行了异常检测的仿真实验。实验结果表明：在异常检测中采用IIGA优化的隶属函数后，异常检测的检测率提高。

参考文献

[1] Zhu TianQing, Xiong Ping. Optimization Of Membership Functions In Anomaly Detection Based On Fuzzy Data Mining[C]. Guangzhou: The Fourth International Conference on Machine Learning and Cybernetics, 2005: 1987—1992.

[2] Bridges SM, Vaughn RB. Fuzzy Data Mining And Genetic Algorithms Applied To Intrusion Detection[C]. Canda: 12th Annual Canadian Information Technology Security Symposium, 2000: 109—122.

[3] 王永杰，鲜明. 模糊逻辑和遗传算法在IIDS中的应用[J]. 计算机工程，2004，30(9)：134—135.

[4] 杨剑峰. 基于遗传算法和蚂蚁算法求解函数优化问题[J]. 浙江大学学报，2007，41(3)：427—430.

[5] 高岩. 免疫遗传算法的研究及其在函数优化中的应用[J]. 微计算机信息，2007，23(23)：183—184.

An Improved Immune Genetic Algorithm and Its Application in Anomaly Detection

LIU Dong, ZHENG Ning, XU Hai-tao

(*School of Computer, Hangzhou Dianzi University, Hangzhou Zhejiang* 310018, *China*)

Abstract: In view of immune genetic algorithm's shortages, an improved algorithm was introduced. The proposed algorithm used isolated niche technology, improved the cross and mutation operation based on immune genetic algorithm. The optimized membership functions were used in fuzzy association rules mining to anomaly detection. It could magnify the similarity between normal association rule sets, and reduce the similarity between a normal and an abnormal association rule set at the same time. So it could improve the performance of anomaly detection. Feasibility of the algorithm was verified by experiments on anomaly detection based on network traffic.

Key words: anomaly detection; membership functions; immune genetic algorithm; isolated niche