# 网络异常检测中一种隶属度函数优化的新方法

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

**摘 要**:对遗传算法的特性进行了研究,在其基础上引入了隔离小生境技术和自适应机制,提出一种改进算法.改进后的算法更接近于实际的进化行为,能够提高优化性能.以网络流量为数据的异常检测实验仿真表明,改进的算法比传统的遗传算法更有效,能够提高异常检测的性能.

**关键词**:异常检测;隶属度函数;遗传算法;隔离小生境;自适应机制
**中图分类号**:TP301.6 **文献标识码**:A **文章编号**:1000-7180(2008)06-0186-04

# A New Method of Optimizing Membership Functions in Net Anomaly Detection

LIU Dong, ZHENG Ning, XU Hai-tao, YANG Jie
(College of Computer, Hangzhou Dianzi University, Hangzhou 310018, China)

**Abstract**: An improved algorithm was introduced, it used isolated niche technique and adaptive mechanism based on genetic algorithm. The improved algorithm could simulate the fact evolution behaviors better, so it could enhance the optimized performance. Experiments on anomaly detection to network traffic prove that the improved algorithm is more efficient than common genetic algorithm, and it can improve the capability of anomaly detection.

**Key words**: anomaly detection; membership functions; genetic algorithm; isolated niche; adaptive mechanism

## 1 引言

如何优化隶属度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.这是因为隶属度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)对隶属度函数进行优化,并取得了一定的效果,提高了异常检测的效率[1-2].但传统的遗传算法本身存在着许多不足,例如在解分布不均匀时易出现未成熟收敛,以及容易陷入局部最优等.

针对上述情况,文中提出一种改进遗传算法——自适应隔离小生境遗传算法(Adaptive Isolated Niche Genetic Algorithm, AINGA).该算法在遗传算法的基础上引入了隔离小生境技术,提高了算法对多峰值函数的优化能力;同时用自适应机制改进变异算子,以提高局部寻优能力.将该算法应用到基于模糊关联规则挖掘的异常检测中,可以提高异常检测的检测率.

## 2 基于模糊关联规则挖掘的异常检测

### 2.1 模糊关联规则

关联规则是形如 $X \rightarrow Y, c, s$ 的蕴涵式,其中 $X = \{x, x_2, \cdots, x_p\}$ 和 $Y = \{y_1, y_2, \cdots, y_q\}$ 是表属性集的子集,且 $X \cap Y = \Phi$;$s$ 和 $c$ 分别是支持度和置信度[3].模糊关联规则是用由多个属性组成的模糊集合 $w_{\text{fuzzy}} = \{w_1, w_2, \cdots, w_p\}$ 和相应的隶属度函数集 $F_w = \{f_{w1}, f_{w2}, \cdots, f_{wp}\}$ 取代特定的连续属性 $w$.一个数据 $d$ 对某个模糊属性 $w_i$($w_i \in w_{\text{fuzzy}}, 1 \leqslant i \leqslant p$)的支持记数是通过该数据对于该属性的隶

**收稿日期**:2007-08-10
**基金项目**:浙江省自然科学基金项目(Y106176);浙江省科技厅科技计划项目(2007C33058)

?1994-2016 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

属度 vote($d$) = $f_{wi}(d)$ 来表示的.则对于模糊属性的关联规则转化为:$< X, A > \rightarrow < Y, B > [c, s]$. 其中,$A = \{w_{x1}, w_{x2}, \cdots, w_{xp}\}$ 和 $B = \{w_{y1}, w_{y2}, \cdots, w_{yq}\}$ 分别是与 $X$ 和 $Y$ 相关联的模糊集.

此外,对于隶属度函数的选取,文中采用异常检测中常用的3个标准隶属度函数 S、PI、Z,将模糊变量分为 High、Medium、Low3 个模糊集合.这 3 个函数的描述如图 1 所示.

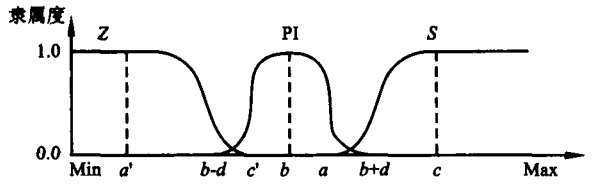


图 1　隶属度函数

## 2.2　在异常检测中运用模糊关联规则挖掘

模糊关联规则挖掘算法在异常检测中应用的步骤如下.首先在系统正常状态下挖掘模糊关联规则集 $S_1$.其次,挖掘系统当前状态下的模糊关联规则集 $S_2$.在得到以上两个规则集后,引入相似度的概念来表征两个规则集之间的相似程度.在异常检测中通常设置一个阈值,若两规则集之间的相似度小于阈值,则表示当前系统状态是异常的;否则,则当前系统状态是正常的.

对于相似度的计算,采用如下方法.给定两条关联规则 $R_1: X \rightarrow Y, c, s$, $R_2: X' \rightarrow Y', c', s'$,则 $R_1, R_2$ 之间的相似度:当 $X = X'$ 且 $Y = Y'$ 时,

$$\text{Similarity}(R_1, R_2) = \text{Max}\left(0, 1 - \text{Max}\left(\frac{|c - c'|}{c}, \frac{|s - s'|}{s}\right)\right),$$

否则 $\text{Similarity}(R_1, R_2) = 0$.

关联规则集 $S_1$ 与 $S_2$ 之间的相似度:

$$\text{Similarity}(S_1, S_2) = \frac{s}{|S_1|}\frac{s}{|S_2|},$$

式中,$|S_1|$ 和 $|S_2|$ 分别是关联规则集和所对应的规则数量,

$$s = \sum_{\substack{\forall R_1 \in S_1 \\ \forall R_2 \in S_2}} \text{Similarity}(R_1, R_2).$$

# 3　改进遗传算法(AINGA)

GA 是模拟生物进化过程的计算模型.它作为一种新的全局优化搜索算法,具有简单通用、鲁棒性强、适于并行处理的特点.但简单遗传算法也存在着许多不足,例如在解分布不均匀时易出现未成熟收敛,以及容易陷入局部最优等[4].

为了使遗传算法更接近于实际的进化行为,文中借鉴了生物界"隔离"遗传的思想,把隔离小生境技术引入到遗传算法中,以提高算法对多峰值函数的优化能力.同时用自适应策略改进变异算子,以增加个体多样性,提高算法的局部寻优能力.下面分别加以具体介绍.

## 3.1　隔离小生境技术

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,小种群沿着不同的方向进化,渐渐由同一物种分化为不同的物种.在种群分化过程中,隔离起着非常重要的作用[5].为了保持物种的多样性,文中设计的算法中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 $k$ 第 $t+1$ 代的规模 $n_{k+1}(t+1)$ 按照下式确定:

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

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

在进化过程中,有些种群因缺乏竞争将面临灭种的局面,为了保持种群的多样性,需要对种群规模进行一定的干预,本算法限制种群的最大规模不可以大于 $M_{\max}$,种群的最小规模不可以小于 $M_{\min}$.即种群规模要满足式(2):

$$M_{\min} \leqslant n_k(t+1) \leqslant M_{\max} \tag{2}$$

## 3.2　自适应机制

自适应机制的主要思想是将算法过程中的中间信息反馈到算法中去,对迭代过程进行跟踪,根据进化的实际方向实时改变策略,以提高算法与对象的耦合程度.变异算子在整个进化过程中十分重要,它们直接影响到算法的性能.文中设计的变异算子能够随进化过程进行自适应调节.

对于变异算子,使变异率具有自适应性.在正常的进化过程中将变异率维持在一个较低的水平,让其快速进化;未进化时间越长,则依靠现有群体找到最优解的可能性越小,需要扩大变异率.文中定义的变异率为

$$P_m(g) = P_m(\min) + N_G a_c \tag{3}$$

式中,$g$ 表示当前代数,$P_m(\min)$ 表示最小变异率,$a_c$ 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.

?1994-2016 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

## 4 AINGA优化隶属度函数

以文中所选隶属度函数为例,用AINGA算法对其进行优化,以期搜索到最佳的隶属度函数,从而提高异常检测的性能.

(1) 编码表示

采用实数编码来表示遗传基因的染色体,这不仅能够减小存储容量,而且可以提高算法精度.对于隶属度函数 $S$、PI、$Z$,每个函数都需要确定2个参数,则定义模糊变量的3个函数总共需要6个参数,则染色体的数据结构如图2所示.

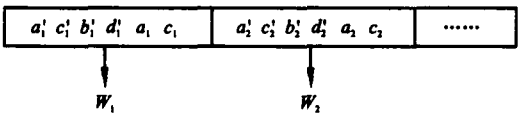


图2 染色体的数据结构

(2) 种群初始化

随机产生 $M$ 个个体.而一个个体的产生按以下步骤操作.针对一个模糊变量 $W$,随机产生关于它的隶属度函数的6个参数 $a'$,$c'$,$b$,$d$,$a$,$c$ 的值.另外,由图1可以看出,隶属度函数的6个参数需满足式(4).

$$\min < a' < b - d < c' < b < a < b + d < c < \max \tag{4}$$

式中,min、max是该模糊变量的最小值与最大值.对这6个参数进行编码,就构成了染色体中的一段基因;若有多个模糊变量,则对每个模糊变量进行相同的操作而得到基因片段,最后将各基因片段连接起来就形成一个个体.

(3) 初始种群的隔离

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

针对文中研究内容,设计的适应度函数是建立在模糊关联规则相似度基础之上的.隶属度函数优化的目标是尽量使得用户正常状态和参考状态关联规则集之间的相似度最大化,同时使得用户异常状态与参考状态关联规则集之间的相似度最小化.

采用两组审计数据来进行测试:一组是没有入侵的正常数据,把这组数据再分成两部分,分别用reference和normal标识;另一组是含有入侵的异常数据abnormal.reference为训练数据,用于挖掘正常状态参考关联规则集 $S_r$;normal为正常数据,用于挖掘正常状态关联规则集 $S_n$;abnormal用来挖掘异常状态关联规则集 $S_a$.此外,对于那些不符合式(4)要求的个体,是不具有实际意义的,对于这些个体应使其个体适应度为0.适应度函数的定义为

$$F = \begin{cases} S_{rn}/(1 + S_{ra}), \text{满足式(4)} \\ 0, \text{其他情况} \end{cases} \tag{5}$$

式中,$S_{rn}$ = Similarity($S_r$,$S_n$),

$S_{ra}$ = Similarity($S_r$,$S_a$).

(5) 子种群规模确定

按式(1)和式(2)计算下代子种群规模.

(6) 选择操作

采用轮盘赌选择算法选择下一代个体.在进化过程中为了防止最优个体丢失,让上一代的最优个体不参加遗传直接保存至下一代.

(7) 交叉操作

选择要进行交叉的两个个体,然后用单点交叉方式进行交叉.

(8) 变异操作

判断染色体上的每个基因是否发生变异.对于要发生变异的基因,将其对应的参数值在其取值范围内做小幅度的变化.

(9) 算法终止判断

优化的目标是寻找最优的隶属度函数,而最优解的适应度值是不确定的,因此通过限定最大遗传代数来终止算法.若满足条件则结束算法,否则继续执行步骤(4)~(8).

## 5 实验仿真

### 5.1 实验数据属性和模糊集合

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;net1数据用abnormal标识.

为了得到实验所需的模糊集合,对原始的网络流量数据进行处理.得到TS、TF和TR3组数值属性,它们分别代表过去两秒内,包头中包含同步标志SYN、连接结束标志FIN和复位标志RST的TCP包的数量.这3个数值属性可以被看作是模糊变量.用3个标准函数 $S$、PI、$Z$,分别将这3个模糊变量分为High、Medium、Low3个模糊集合.

### 5.2 参数优化及检测

数据包含3个模糊变量,其隶属度函数由6个参数确定,则参数集包含18个参数.随机形成40个个体组成初始群体,把初始群体平均分成两组,形成

?1994-2016 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

初始的隔离子种群.对初始群体中的每个个体,取出各参数对应的值.取最小支持度为0.1,最小置信度为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 $S_{rn}$ 和 $S_{ra}$ 以及每个个体的适应度.取自适应变异率 $p_m(g)=0.04+0.01\times N_G$,交叉概率 $p_c=0.5$,迭代数为50代,按照AINGA算法进行遗传操作,得到最优个体.取相同的参数,用传统遗传算法进行与上述同样的优化实验.

在种群规模分别为20、100个个体,相应的遗传代数分别为30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所对应的隶属度函数进行模糊关联规则挖掘,并计算 $S_{rn}$ 和 $S_{ra}$,实验结果如表1所示.

表1　AINGA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 |
|---|---|---|---|---|
| AINGA优化后 | $S_{rn}$ | 0.931 7 | 0.968 4 | 0.970 6 |
| | $S_{ra}$ | 0.129 1 | 0.094 7 | 0 |
| GA优化后 | $S_{rn}$ | 0.789 9 | 0.862 1 | 0.875 1 |
| | $S_{ra}$ | 0.144 3 | 0.106 7 | 0.053 6 |

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AINGA优化的隶属度函数后,正常关联规则集之间的相似度($S_{rn}$)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,提高了异常检测的检测率.

## 6　结束语

文中在遗传算法的基础上,引入隔离小生境技术,使用自适应机制改进了变异算子,提出一种改进算法.将该算法应用到基于模糊关联规则挖掘的异常检测中,能够提高正常关联规则集之间的相似度,降低正常与异常关联规则集之间的相似度,从而使异常检测的检测率提高.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,该算法是可行的、有效的.

**参考文献:**

[1] Zhu Tian-Qing, Xiong Ping. Optimization of membership functions in anomaly detection based on fuzzy data mining[C]// The Fourth International Conference on Machine Learning and Cybernetics. Guangzhou, 2005:1987-1992.

[2] 蔡伟鸿,刘震,王美林.基于模糊逻辑和免疫遗传算法的入侵检测[J].计算机工程,2006,32(7):151-153.

[3] 金榜.一种数据挖掘算法在入侵检测系统中的应用[J].微电子学与计算机,2006,23(1):181-183.

[4] 汤亚玲,崔志明.遗传算法在Web关联挖掘中的应用[J].微电子学与计算机,2005,22(10):4-6.

[5] 林焰.隔离小生境遗传算法研究[J].系统工程学报,2000,15(1):86-91.

**作者简介:**

**刘　董**　女,(1982-),硕士研究生.研究方向为网络安全.

**郑　宁**　男,(1961-),博士生导师.研究方向为计算机应用.

**徐海涛**　男,(1978-),硕士.研究方向为数据仓库和信息处理.

**杨　洁**　男,(1984-),硕士研究生.研究方向为网络安全.

(上接第185页)

[2] Art Baker, Jerry Lozano. Windows 2000设备驱动程序设计指南[M].施诺,译.北京:机械工业出版社,2001:113-129.

[3] 雷艳静,魏建军,王珂,等.面向机群系统的高速光纤传输网络接口卡设计[J].计算机工程与应用,2006(13):19-21.

[4] 郑魁,杨志义,曾小芸.基于Linux的光纤通道网卡驱动程序开发[J].微电子学与计算机,2007,24(1):88-90.

[5] 贾阳春,刘小丹.Windows 2000下应用WDM驱动程序实现I/O端口直接访问[J].辽宁师范大学学报,2003,26(4):385-388.

[6] 薛纪文,王会燃,加云岗.基于WDM的I/O端口访问设计及实现[J].微电子学与计算机,2005,22(6):91-93.

**作者简介:**

**郭海山**　男,(1979-),硕士研究生.研究方向为高性能计算与群机系统.

**冯　萍**　女,(1955-),教授.研究方向为高性能计算、计算机应用.

**许　伟**　男,(1982-),硕士研究生.研究方向为高性能计算与群机系统.

?1994-2016 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net