文章编号: 1001-9081(2006)01-0210-03

基于模糊数据挖掘与遗传算法的异常检测方法

孙 东[1], 黄天戍[1], 秦丙栓[2], 朱天清[3]
(1. 武汉大学 电子信息学院, 湖北 武汉 430079;
2. 中国有线电视网络公司 大客户部, 北京 100053;
3. 武汉工业学院 计算机与信息工程系, 湖北 武汉 430023)
联eastsun11@263.net贩

摘 要: 建立合适的隶属度函数是入侵检测中应用模糊数据挖掘所面临的一个难点。针对这一问题, 提出了在异常检测中运用遗传算法对隶属度函数的参数进行优化的方法。将隶属度函数的参数组合成有序的参数集并编码为遗传个体, 在个体的遗传进化中嵌入模糊数据挖掘, 可以搜索到最佳的参数集。采用这一参数集, 能够在实时检测中最大限度地将系统正常状态与异常状态区分开来, 提高异常检测的准确性。最后, 对网络流量的异常检测实验验证了这一方法的可行性。

关键词: 异常检测; 模糊数据挖掘; 遗传算法
中图分类号: TP311.13 文献标识码: A

Anomaly detection approach based on fuzzy data mining and genetic algorithm

SUN Dong[1], HUANG Tian-shu[1], QIN Bing-shuan[2], ZHU Tian-qing[3]
(1. *School of Electron and Information, Wuhan University, Wuhan Hubei 430079, China*;
2. *Major Accounts Department, China Cable Television Network Co., Beijing 100053, China*;
3. *Department of Computer & Information Engineering, Wuhan Industry Institute, Wuhan Hubei 430023, China*)

Abstract: Defining appropriate membership functions is a difficult task in fuzzy data mining to detect intrusions. To solve the problem, an approach that applies genetic algorithm to optimize parameters of membership functions in anomaly detection was presented. Parameters of membership functions were arranged into a sequential parameter set coded to an individual. An optimal parameter set could be derived by embedding fuzzy data mining in the process of evolution of individual. With the parameter set in anomaly detection, normal state of protected system could be differentiated from anomalous state in the most extent, and the veracity of anomaly detection was improved greatly. Experiments on anomaly detection to network traffic prove the feasibility of the approach.

Key words: anomaly detection; fuzzy data mining; genetic algorithm

入侵检测系统(IDS)是信息安全体系的重要组成部分, 它通过区分系统的正常行为及异常行为来发现入侵。目前入侵检测的方法主要有两种, 即滥用检测和异常检测[1]。其中, 异常检测由于能够检测到未知及新型攻击, 是一种更为严格的入侵检测方法。

异常检测是基于这样一个假设, 即入侵者的活动在某种程度上与正常用户的行为有所不同。其总体思想是, 首先用网络及系统的某些特性参数和阈值来定义正常用户行为和系统正常轮廓, 然后用这一正常轮廓(profile)与系统的暂态轮廓进行对比, 若有超出某种程度的差异, 则确定为异常。因此, 检测异常和入侵实质上就是检测这些特性参数与正常状态值的背离程度[2]。

异常检测的主要难点在于如何全面地定义系统的正常轮廓。针对这一问题, Wenke Lee等提出了在入侵检测中应用数据挖掘技术[3,4]。另外, 针对数据挖掘中的"尖锐边界"问题及系统安全本身的模糊性, Susan M. 等在数据挖掘中引入了模糊集合论, 即模糊数据挖掘, 并应用到入侵检测中[5]。然而模糊集合亦有其局限性, 即模糊集合隶属度函数及其参数的确定过于依赖于专家领域知识, 影响了入侵检测的准确性。

本文在模糊数据挖掘中引入遗传优化算法, 提出了基于模糊数据挖掘与遗传优化的异常检测方法, 通过对模糊集合隶属度函数的各参数进行组合并遗传优化, 以搜索最佳的参数组合。最后通过对网络流量的异常检测实验, 验证了这一方法的有效性。

1 模糊数据挖掘

由于数据挖掘通常只能对离散值进行处理, 在数据预处理中要将连续属性域划分为若干离散区间, 这就导致了所谓的"尖锐边界"问题[6]。模糊数据挖掘(在数据挖掘中引入模糊集合)的应用即是为了解决这一问题。

1.1 模糊关联规则挖掘

关联规则挖掘是入侵检测中普遍应用的数据挖掘方法, 用于发现数据库表中各属性间隐含的关系, 以形成关联规则: $X \rightarrow Y\ [s, c]$, 其中, $X = \{x_1, x_2, \cdots, x_p\}$ 和 $Y = \{y_1, y_2, \cdots, y_q\}$ 是表属性集的子集, 且 $X \cap Y = \varnothing$, s和 c分别为支持度和置信

收稿日期: 2005-07-22; 修订日期: 2005-10-10 基金项目: 公安部科研基金资助项目(200342 823 01)

作者简介: 孙东, 博士研究生, 主要研究方向: 网络安全; 黄天戍, 教授, 主要研究方向: 通信与网络安全; 秦丙栓, 助工, 主要研究方向: 广播电视通信网络; 朱天清, 硕士, 主要研究方向: 网络安全.

度。

一个典型的关联规则挖掘算法是 agrawal和 srikant提出的 Apriori算法[7]。针对应用该算法所导致的“尖锐边界”问题, Kuok Fu等在文献[6]中提出了模糊关联规则的概念, 将特定的连续属性 w取代为由多个属性组成的模糊集合 $w_{fuzzy}=\{w_1, w_2, \cdots, w_p\}$和相应的隶属度函数集 $F_w=\{f_{w1}, f_{w2}, \cdots, f_{wp}\}$。一个数据 d对各属性 $w_i \in w_{fuzzy}$ ($0 \leqslant i \leqslant p$) 的支持记数则由该数据对于该属性的隶属度 $vote(d)=f_{wi}(d)$来表示。相应地, 关联规则的形式转化为:

$$<X, A>\rightarrow<Y, B>\ [c, s] \qquad (1)$$

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。显然, 模糊集合的引入, 扩展了数据库表的属性, 增加了关联规则挖掘的计算量。

另外, 模糊集合隶属度函数的确定是一个十分重要的环节, 它是定量地分析模糊属性的基础。但隶属度函数的确定具有一定的主观性, 取决于人们对模糊集合的认识程度、实践经验, 甚至包含一定程度的心理状态。常见的模糊分布包括偏小型 (Z型函数)、偏大型 (S型函数) 和对称型 (π型函数)。一个标准的隶属度函数 F如图 1所示。

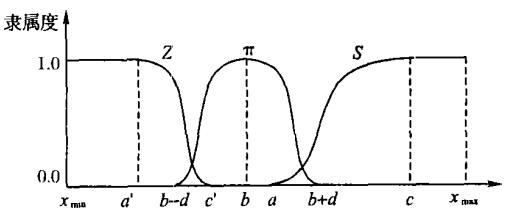

图 1　模糊集合隶属度函数 F

图 1中, 隶属度函数 F由 $Z(low)$、$S(high)$和 $\pi(medium)$ 3个函数段组成, 每个函数段分别由 2个参数确定。在基于异常检测的入侵检测系统研究中普遍应用了该函数。

1.2　在异常检测中应用模糊关联规则挖掘

在异常检测中应用模糊关联规则挖掘首先需要形成挖掘的对象, 即系统状态数据库。数据库的属性是系统用户层、系统层、进程层和数据链路层的各种参数。如用户层参数包括用户类型、用户权限、登录 /退出的时间和地点、访问的资源等; 数据链路层参数包括连接数量和状态、连接时间、使用的协议和端口等。文献[8]描述了各层所要监测的参数和得到这些参数值的有关系统命令。

在系统正常状态下, 统计各参数在多个时间域中的值, 构成系统正常状态数据库, 应用隶属度函数 F对数据库的连续属性进行模糊处理, 然后通过模糊关联规则挖掘, 建立系统正常状态下的关联规则集 S, 代表系统的正常轮廓。在检测时, 挖掘系统在暂态模式下的关联规则集 S_1, 通过计算两规则集的相似度[9] $similarity(S, S_1)$, 即用相似度来表示系统当前状态与正常状态的背离程度, 以确定系统是否处于异常状态。

对于相似度的计算, 我们采用如下方法。

给定两个关联规则:

$R_1: X\rightarrow Y, c, s$; $R_2: X'\rightarrow Y', c', s'$, 如果 $X=X'$且 $Y=Y'$, 则这两个关联规则的相似度为:

$$similarity(R_1, R_2)=\max\left(0, 1-\max\left(\frac{|c-c'|}{c}, \frac{|s-s'|}{s}\right)\right) \qquad (2)$$

若 $X\neq X'$或 $Y\neq Y'$, 则 $similarity(R_1, R_2)=0$。

两个关联规则集 S_1和 S_2的相似度为:

$$similarity(S_1, S_2)=\frac{S^2}{|S_1|*|S_2|} \qquad (3)$$

其中, $s=\sum_{\substack{\forall R_1\in S_2\\ \forall R_2\in S_2}} similarity(R_1, R_2)$, $|S_1|$和 $|S_2|$分别为 S_1和 S_2的规则数量。

显然, $0\leqslant similarity(S_1, S_2)\leqslant 1$。如果 $similarity(S_1, S_2)$小于预先设定的阈值, 则表示系统当前状态与正常状态的背离程度较大, 处于异常状态。

2　遗传算法的应用

2.1　遗传算法

遗传算法是模仿生物遗传与进化过程而得出的一种随机优化方法[10]。应用遗传算法时, 用一个个体代表一个可能的解, 而个体的性能优良与否由适应度函数衡量。适应度函数通过测试个体染色体是否满足算法设计者的要求来衡量个体性能。在达尔文进化论中, 低性能的个体将会从群体中去除, 而高性能的个体被复制、变异, 取代被去除的个体。与生物变异相似, 一些进行随机变异的个体在理论上性能得以提高, 直至达到最好的适应值或适应值误差范围, 即找到理想的个体。若理想个体没有找到, 遗传算法在预定义的代数到达最大值时结束。

其主要步骤包括:

1)对研究的变量进行编码, 并随机地建立一个初始群体;

2)计算群体中诸个体的适应度;

3)执行产生新群体的操作, 包括复制、交换、突变等;

4)根据某种条件判断计算过程是否结束, 不满足则返回步骤 2重复执行。

2.2　异常检测中隶属度函数的参数优化

在异常检测中应用模糊集合时, 由于隶属度函数的确定过于依赖于专家领域知识, 具有相当的主观性。这里我们以图 1所示的隶属度函数 F为例, 应用遗传算法对其参数进行遗传优化, 以搜索最佳的隶属度函数。

2.2.1　编码

隶属度函数 F包含 6个参数, 即要对 6个参数分别编码, 并连接起来形成一个由“0”和“1”组成的参数组。当数据库中包含多个模糊属性时, 则编码为多个参数组, 连接后成为一个个体。例如, 当存在 2个模糊属性时, 个体是一个对 12个参数编码后的 0/1字符串, 如图 2所示。

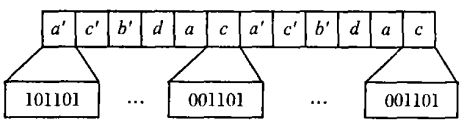

图 2　隶属度函数 F的参数编码

编码需要首先确定变元的取值区间 $[x_{min}, x_{max}]$。显然, x_{min}和 x_{max}分别是数据库中模糊属性对应的一列数据中的最小值和最大值。一般来说, 变元的取值区间越大, 则表示变元的编码字符串越长, 计算量也将随字符串位数的增长呈指数倍增长。为此, 我们可根据隶属度函数中各参数间的关系, 有目的地对变元的取值区间加以限制。考察函数 F中各参数间的关系, 可知它们须满足不等式:

$$a'<b-d<c'<b<a<b+d<c \qquad (4)$$

根据它们的大致相对位置确定各参数的范围, 使取值区

间长度为(x_{min} - x_{max})/2缩短了字符串长度, 如:

$x_{min} \leqslant a' \leqslant (x_{min} + x_{max})/2$;

$(7x_{min} + x_{max})/8 \leqslant b \leqslant (3x_{min} + 5x_{max})/8$等。

对于每个参数, 根据(x_{min} - x_{max})/2的大小, 确定字符串长 N, 确定 2^N 个可行解并分别编码, 最后组合成一个个体, 即染色体。按照同样的方法随机生成多个个体组成初始群体。

2.2.2 个体的适应度

适应度是描述群体中个体优劣的尺度。

在异常检测中, 最佳的隶属度函数应该使系统正常状态关联规则集 S_1 与异常状态关联规则集 S_2 的相似度为最小。

设 $P = \{x_1, x_2, \cdots, x_n\}$ 是所有隶属度函数的参数组成的集合, 其中 n 为参数个数。一个个体即表示一个参数集。可取适应度函数为 $f(P) = 1 - similarity(P, S_1, S_2)$, 即隶属度函数的各参数取该个体所表示的值时 S_1 与 S_2 的相似度与最大值 1的差值, 其中 S_1、S_2 分别是参数集为 P 时通过模糊关联规则挖掘产生的系统正常状态和异常状态规则集。

另外, 隶属度函数的参数须满足不等式(4), 不满足该约束的个体属于不可行解, 因此最终确定适应度函数为:

$$f(P) = \begin{cases} 1 - similarity(P, S_1, S_2), & P\text{满足式(4)} \\ 0 & \text{其他} \end{cases} \quad (5)$$

2.2.3 遗传算子

遗传算法中最主要的算子是复制、交换和突变。

取复制概率 P_d 为 10% ~ 20%, 采用轮盘方式随机选择复制对象, 或者直接选取适应度最小的个体, 复制为适应度最大的个体。同时, 由于遗传算法收敛于最优解的概率小于 1, 因此必须实施最优保存策略, 在每代个体中保留上一代的最优个体。

设群体中个体总数为 n, 交换概率为 P_c(一般为 0.5 ~ 0.8), 则交换的个体数目为 $n \cdot P_c$。从群体中随机选取被交换的个体实行交换操作。另外, 交换点的选择也是随机的。在异常检测的模糊关联规则挖掘中, 由于数据库中往往包含有多个连续属性, 在模糊处理时则包含多个隶属度函数, 参数个数较多, 因此个体的字符串长度较长, 在交换时应采用两点或多点交换。

突变概率 P_m 一般很小, 为 0.001 ~ 0.01。设个体长度为 L, 则突变字符的个数为 $n \cdot P_m \cdot L$。在所有个体中按均匀分布选择突变字符, 并对突变字符作补运算后取代原个体。

2.2.4 运算终止准则

在异常检测中应用遗传算法, 是为了寻求最优的隶属度函数参数集, 其最优解的适应度值是不确定的, 在定义运算终止准则时, 可采取两种方法。其一, 依据专家经验或对适应度的期望确定一个理想适应度值, 一旦某代最优个体的适应度超过了这个理想值, 则运算停止; 其二, 直接规定迭代次数, 达到这个次数时即停止运算, 选取适应度最大的解。

3 网络流量的异常检测实验

运用以上阐述的方法, 我们开发了一套综合应用模糊关联规则挖掘和遗传算法的软件——FGS系统, 对实验环境下局域网的流量进行分析, 验证这一方法的可行性。

3.1 网络流量的属性及其模糊集合

根据局域网的实际情况, 我们选择与网络流量相关的四个属性来对系统进行分析: TCP和 UDP包在全部数据包中的比例 P_{tcp} 和 P_{udp}、网络中每秒的平均数据包数量 *Avg packet/sec* 以及每秒平均数据位 *Avg Mbit/sec* 分别在正常状态和异常状态(对网络中的某台主机作拒绝服务攻击)下收集一组数据, 每 2分钟取一次样, 各得到 10组数据, 如表 1所示。

表 1 实验数据

P_{tcp}	P_{udp}	Avg. pkt/sec	Avg. Mb/sec	状态
94.5	0.6	169.368	0.527	正常
94.5	0.7	171.532	0.548	正常
爛	爛	爛	爛	爛
90.2	0.4	143.566	0.416	正常
96.2	1.0	181.477	0.523	正常
95.2	0.3	169.542	0.530	异常
95.9	0.4	171.837	0.531	异常
爛	爛	爛	爛	爛
96.2	0.9	171.476	0.523	异常
93.8	0.4	183.937	0.596	异常

将以上 4个属性划分为 3个模糊集合, 即 *low*、*medium* 和 *high*, 隶属度函数 F 为:

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

$$Z(x, a', c') = 1 - S(x, a, c)$$

$$\pi(x, a, c) = \begin{cases} S(x, b-d, b), & x \leqslant b \\ Z(x, b, b+d), & x > b \end{cases}$$

3.2 参数优化

表 1中, 数据库包含 4个连续属性, 每个属性分为 3个模糊集合, 其隶属度函数由 6个参数确定, 因此整个参数集包含 24个参数。根据表 1中的各属性所对应数据的数值范围, 我们用 5位的 0/1字符串来表示 P_{tcp} 和 *Avg packet/sec* 的隶属度函数中的参数, 用 3位的 0/1字符串来表示 P_{udp} 和 *Avg Mbit/sec* 的隶属度函数中的参数。因此一个个体的长度为:

$L = 5\times 6 + 3\times 6 + 5\times 6 + 3\times 6 = 96$(位)

按照 2.2.1节中阐述的方法进行编码, 随机形成 5个个体, 组成初始群体。

对初始群体中的每个个体, 取出各参数相应的字符串, 将它们转化为实数, 得到各参数的数值。取最小支持度 $minsupport = 0.25$, 最小置信度 $minconfidence = 0.6$, 按照模糊关联规则挖掘的方法分别从正常状态数据和异常状态数据中挖掘出关联规则集 S_1 和 S_2, 计算它们的相似度, 根据式(5)得到全部个体的适应度, 如表 2所示。

将适应度最小的个体复制为适应度最大的个体, 其他 4个个体采用 3点交换方式两两交换, 交换点随机选择, 交换字符串长度为 24位。取突变概率 $P_m = 0.05$, 在 4个个体中各随机选取一位作补运算, 最后得到下一代个体。

表 2 个体的适应度和相似度

序号	个体	相似度	适应度
1	0000爛01100	0.712	0.288
2	00100爛10110	0.848	0.152
3	10000爛01101	0.741	0.259
4	00010爛01110	0.587	0.413
5	00101爛11110	0.652	0.348

取迭代数为 1000代, 重复以上的步骤, 直到循环结束, 得到最优个体。

(下转第 215页)

DD International Conference on Knowledge Discovery in Database and Data mining[C]. Edmonton,Alberta,Canada,2002. 217 - 228.

[3] RIZVI S,HARITSA JR. Privacy-preserving association rule mining [C]. Proceedings of 28th International Conference on Very large Data Bases. VLDB[C],2002.

[4] LINDELL Y,PINKAS B. Privacy preserving data mining[C]. In Advances in Cryptology - CRYPTO 2000[C]. Springer Verlag,2000. 36 - 54.

[5] OLIVEIRA SRM,ZAïANE OR. Privacy preserving frequent item set mining[C]. Proceedings of the IEEE ICDM Workshop on Privacy,Security and Data Mining[C]. Maebashi City,Japan,2002. 43 - 54.

[6] RIZVI S,HARITSA JR. Maintaining data privacy in association rule mining[C]. Proceedings of the 28th International Conference on Very Large Databases[C],2002.

[7] CLIFTON C,KANTARCIOGLOU M,LIN X,et al. Tools for privacy preserving distributed data mining[C]. SIGKDD Explorations,2002,[C],4(2):28 - 34.

[8] AGRAWAL D,AGGARWAL CC. On the design and quantification of privacy preserving data mining algorithms[C]. Proceedings of the 20th ACM Symposium on Principles of Database Systems[C],2001. 247 - 255.

(上接第 212页)

迭代中各代最优个体的适应度和规则集相似度变化如图 3所示。

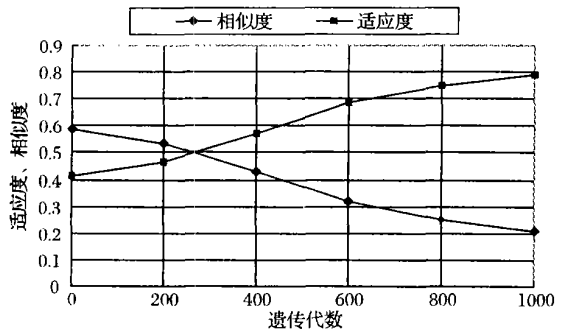

图 3　各代最优个体的适应度和相似度

由图 3可以看出,遗传 1000代后得到的最优个体(适应度 =0.789)使得正常状态规则集与异常状态规则集的相似度(=0.211)最小。

在多个异常状态下收集数据,重复以上的实验过程,结果如图 4所示。

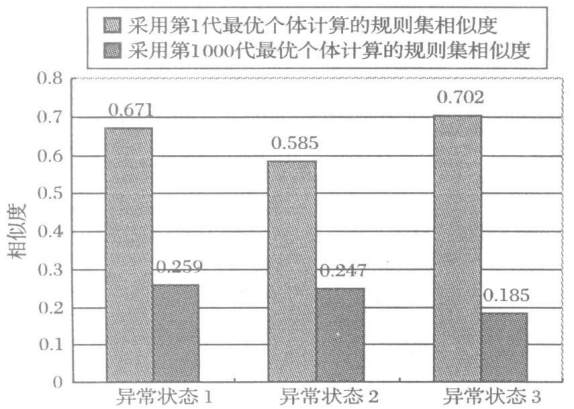

图 4　遗传优化前、后的相似度比较

实验结果表明,通过遗传优化能够有效地搜索到一个理想的隶属度函数参数集,采用这个参数集并通过模糊关联规则挖掘,大大降低了异常状态与正常状态下规则集的相似度,最大限度地将系统正常状态与异常状态区分开来。与由专家领域知识来确定参数集相比,该方法更科学,性能更优异。

4　结语

在异常检测中应用遗传算法,主要用于在入侵检测系统设计中确定较优的隶属度函数的参数,其优点是参数集的初始群体可随机确定,大大减小了对专家领域知识的依赖程度。另外,对于庞大的统计数据而言,应用遗传算法将是一个繁杂、耗时的过程,但这个过程仅限于系统设计阶段,在系统的实时检测中则直接应用所搜索到的最优参数集,因此对系统的运行效率毫无影响。

当然,在实时检测中,隶属度函数也需要不断地调整,这就是入侵检测系统的自适应问题。信息网络系统在实际运行中其正常状态是不断变化的,因此用于异常检测的系统正常状态的"轮廓",即正常状态规则集也必须随之不断更新,最终还是通过调整隶属度函数的参数来实施的。通过遗传算法或者其他智能方法实现参数集自动适应系统正常状态的变化,将是我们进一步的研究内容。

参考文献:

[1] AXELSSON S. Intrusion detection systems:A survey and taxonomy [C]. Technical Report No 99-15[C],Dept of Computer Engineering[C],Chalmers University of Technology[C],Sweden[C],2000.

[2] DEBAR H[C],DACIER M[C],WEPSPI A. A Revised Taxonomy for Intrusion Detection Systems[C]. Technical Report[C],Computer Science / Mathematics[C],IBM Research[C],Zurich Research Laboratory[C],Switzerland[C],1999.

[3] LEE W[C],STOLFO S[C],MOK KW. Mining audit data to build intrusion detection models[C]. Proceedings of the 4th International Conference on Knowledge Discovery and Data Mining[C]. New York:AAAI Press[C],1998.

[4] LEE W[C],STOLFO S[C],CHAN PK[C],et al. Real Time DataMining based Intrusion Detection[C]. Proceedings of DISCEX II[C]. Anaheim[C],USA[C],2001.

[5] BRIDGES SM[C],VAUGHN RB. Intrusion Detection Via Fuzzy Data Mining[C]. Proc of 12th Annual Canadian Information Technology Security Symposium[C]. Ottawa[C],Canada[C],2000.

[6] KUOK C[C],FU A[C],WONG M. Mining fuzzy association rules in databases[C]. SIGMOD Record[C],1998[C],27[C]:41 - 46.

[7] AGRAWAL R[C],SRIKANT R. Fast algorithms for mining association rules[C]. Proceedings of the 20th international conference on very large databases[C]. Santiago[C],Chile[C],1994.

[8] DASGUPTA D[C],GONZALEZ FA. An Intelligent Decision Support System for Intrusion Detection and Response[C]. MMM-ACNS[C],2001.

[9] WANG W. Genetic Algorithm Optimization of Membership Functions for Mining Fuzzy Association Rules[C]. International Joint Conference on Information Systems[C],Fuzzy Theory and Technology Conference[C]. Atlantic City[C],2000.

[10] 郭嗣琮[C],陈刚. 信息科学中的软计算方法[C]. 沈阳:东北大学出版社. 2001.