

基于 Rough 集理论的 Self 集的构造和演化算法

符海东, 李春香

(武汉科技大学 计算机学院, 湖北 武汉 430081)

摘要:提出了一种基于 Rough 集理论的 Self 集构造和演化算法。利用 Rough 集约简算法, 对用户的安全访问行为的数据作规范化处理并进行约简, 从中提取有效的最简规则, 降低了安全数据的冗余, 减轻了特征码构造的负担。使用 Rough 集上、下近似集原理, 构造了上、下近似 Self 集, 实现了 Self 的优化和扩展, 有效地解决了 Self 集的自动演化问题。

关键词:Rough 集; Self 集; 约简; 特征码; 上、下近似集

中图分类号:TP393.08

文献标识码:A

文章编号:1673-629X(2007)12-0060-04

Construction and Evolutionary Algorithm of Self Sets Based on Rough Sets Theory

FU Hai-dong, LI Chun-xiang

(Dept. of Computer Science & Technology, Wuhan University of Science & Technology, Wuhan 430081, China)

Abstract: A construction and evolution algorithm of Self sets based on the rough sets theory is presented. Introducing the reduction algorithm of rough sets, the data of safety visitors' activities are standardized and reduced, and then the most effective and simplest rules are extracted. This algorithm reduced the redundancy of safe data and the burden of construction of characteristic codes. According to lower and upper approximations sets of rough sets, lower and upper approximations Self sets are constructed and the optimization and expansion of Self sets is realized, and the problem of automatically evolution of self sets is effectively resolved.

Key words: rough sets; Self sets; reduction; characteristic codes; lower and upper approximations sets

0 引言

计算机安全问题越来越为人们所关注。目前的安全手段大多都是采取口令验证、防火墙等安全屏障, 其主要特点是不停地补安全漏洞。这些被动的手段不能很好地解决计算机的安全问题, 而借鉴生物免疫系统的计算机安全免疫系统, 以其灵活性和智能性成为解决计算机安全问题的有效手段。

计算机免疫学是一门基于生物免疫学、人工免疫, 以及计算机科学等的交叉学科, 主要利用最新的计算机科学技术, 研究有关人工免疫的理论、规则、算法、模型等, 并将这些理论应用于具体的应用系统, 解决实际的应用课题^[1]。

现在, 国内外有很多学者如 Stephan Forrest、梁意文、李涛等人对计算机安全免疫系统进行研究, 并且取得了一定的进展。

1 人工免疫过程中 Self 集的构造

1.1 国内外学者对 Self 集构造算法的研究

美国新墨西哥大学的 Stephan Forrest 教授领导的研究小组首先提出了利用系统的短调序列的方法来确定用户行为是否合法。入侵行为不仅决定于执行系统调用的种类, 还决定于执行系统调用的顺序。如何判定用户的系统短调序列是“自我”还是“非我”是目前的主要研究方向^[2], 大体上分为规则提取和串匹配^[3]2种方法: 规则提取就是从已有的串中提取出某种规则用来识别入侵的方法; 串匹配就是用未知串与制定出已知“自我”或“非我”的串比对以识别入侵的方法^[4], 但它们的应用前提就是正确地建立 Self 集与 Non-Self 集。梁意文教授提出在定义 Self 集中应允许 Self 的合法变形存在, 同时, 也要识别不合法的改变, 称这种定义为识别特征码。在实际的环境中, Self 集与 Non-Self 集都不可能是静态的, 合法的操作和入侵的方式都是不断变化的, 这就需要 Self 集有不断演化的能力, 能够根据访问的情况自动调整^[5]。

1.2 Self 集构造面临的困难

用于 Self 集构造的信息通常是从海量的通信模式

收稿日期: 2007-03-04

基金项目: 湖北省教育重点科研项目基金(2004D006)

作者简介: 符海东(1971-), 男, 湖北武汉人, 博士, 硕士生导师, 研究方向为模糊推理和遗传算法、网络信息安全。

数据、系统安全日志或用户审计记录中提取出来的,由于上述数据集属性众多,数据量庞大,存在较大的数据冗余,不仅加重了数据管理员的负担,更重要的是增加了 Self 集构造过程中特征码构造的难度,不利于 Self 集的演化。因此在 Self 集构造之前,对信息源进行数据预处理是十分必要的。

在已有的 Self 集构造方法中,既有基于用户行为的构造方法^[6],也有基于进程系统调用的构造方法^[7]。然而,不管是哪种方法,都存在两个问题:

● Self 集中特征码存在大量冗余,加重了匹配的困难,且不利于识别器的培养。

● Self 集不具有自动演化的能力。

1.3 文中的核心思想

文中针对 Self 集构造面临的困难,在计算机免疫模型基础上提出了基于 Rough 集理论的 Self 集的构造和演化算法。首先结合 Rough 集数据约简理论,对安全的数据信息进行约简,降低了安全数据的冗余,减轻了特征码构造的负担,提取出有效的最简规则集和信息可信度,将特征属性的取值离散划为 4 个区间,用二进制编码构造特征码,得出最初 Self 集;然后利用 Rough 集上、下近似集原理,构造 Self 集的上近似 Self 集和下近似 Self 集,实现了 Self 集的演化,Self 集具有自动演化更新的能力,为整个免疫系统的构造打下了良好的基础。

2 Self 集的构造算法设计

2.1 基本定义

以下为几个基本定义^[8]:

定义 1 特征码。

特征码是对用户访问系统资源行为的抽象,具体地说,它是在深刻理解网络协议的机制以及它们安全漏洞的基础上,对正常网络通信模式或是对合法用户表现出来的正常行为特征的编码表示。特征码的选取对于一个免疫系统来说是至关重要的,它直接影响到整个免疫系统的性能。

定义 2 知识表达系统。

Rough 集把客观对象世界抽象为一个知识表达系统: $S = \langle U, C \cup D, V, f \rangle$ 。其中 U 为对象的有限集合; $C \cup D$ 为属性的有限集合, C 表示条件属性集, D 表示决策属性集; V 为属性的值域集; $f: U \times C \cup D \rightarrow V$, f 为信息函数,定义对象的属性值。

定义 3 知识约简。

在知识表达系统 $S = \langle U, C \cup D, V, f \rangle$ 中,对于每个属性子集 $B \subseteq C$,定义一个不可分辨的二元关系 $IND(B) = \{(x, y) \in U \mid \forall b \in B, b(x) =$

$b(y)\}$ 。

对知识约简的要求是,既要消除冗余信息,减少计算量,提高识别速度,又要防止因约简造成重要信息的丢失,影响识别的正确性。Rough 集理论的知识约简表现为不丢失信息的前提下,最简单地表示为知识系统的决策属性与条件属性的依赖性与关联度。

对于一个知识表达系统 $S = \langle U, C \cup D, V, f \rangle$,称满足下列两个条件的 $B \subseteq C$ 是条件属性集 C 的约简:

(1) $IND(B, D) = IND(C, D)$;

(2) 不存在 $B' \subset B$, 使 $IND(B', D) = IND(C, D)$ 。

定义 4 等价关系和等价类。

设 R 是集合 A 上的二元关系,如果它是自反的、对称和传递的,则它是 A 上的等价关系。

设 R 是 A 上的一个等价关系,与 A 中的一个元素 a 相关的所有元素的集合被称作 a 的一个等价类,记成 $[a]_R$,形式地, $[a]_R = \{s \mid (a, s) \in R\}$ 。

定义 5 Rough 集的上近似集与下近似集。

设 $X \subseteq U$, R 是 U 的等价关系,则有:

$R_*(X) = \bigcup \{Y \in U/R : Y \subseteq X\}$

$R^*(X) = \bigcup \{Y \in U/R : Y \cap X \neq \emptyset\}$

分别称它们为 X 的 R -下近似和 R -上近似,其中 \emptyset 是空集, Y 是 U 上按等价关系 R 做成的等价类。

下近似被解释为所有那些被包含在 X 里面的等价类的并集;上近似被解释为所有那些与 X 有交集的等价类的并集。

粗糙集理论的基本思想就是在保持原有数据集的分类与决策能力不变的情况下,运用约简的方法把原来数据表中冗余的数据消去,得到一个约简规则表,然后从该表中提取特征属性。

2.2 基于 Rough 集数据约简算法设计

最初的特征码可以通过对已知的正常的访问行为的学习获得,通常它们是从海量的通信模式数据、系统安全日志或用户审计记录中抽取最能表达合法正常特征的属性集合来进行编码。由于上述数据集属性众多,数据量极大,为了得到最简的有效特征属性,这里引入 Rough 集理论中的决策表约简原理来获得最简的有效规则集。

基于 Rough 集理论的观点,规则的获取可看作一个知识表达、提取有用属性、约简属性表达、得到推理规则的过程。文中首先根据专家经验和领域背景知识,确定各个条件属性的重要度^[9],然后利用 Rough 集方法分析数据之间的联系,挖掘安全数据集中的重要特征信息,提取有效的最简规则,进行二进制编码构造

特征码,用于构建 Self 集。基于 Rough 集数据约简算法如下:

(1)对将要进行处理的安全数据集规范化为一个知识表达系统,即信息表。

(2)根据专家经验和领域背景知识决定各个条件属性的重要度,用 $[0,1]$ 之间的实数表示,数值越大说明其对决策结果的影响越大。

(3)消去对象的重复信息(即行的约简),并记录重复信息的数目 m ,作为信息可信度保存在特征码的可信度位上。利用 Rough 集理论分析条件属性和决策属性的依赖关系,结合各个属性的重要度判断并消去那些冗余条件属性,即条件属性约简。

(4)根据约简的知识表达系统,计算各条规则的核值与其可能的规则简化形式,知识规则集合可以表示为: $\{des(C_i) \rightarrow des(D_j) \mid C_i \cap D_j \neq \emptyset\}$ 。其中 $des(C_i)$ 、 $des(D_j)$ 分别为依照条件属性集 C 和决策属性集 D 所划分的等价类 C_i 、 D_j 的描述。

(5)汇总对应的规则,将获得最终的规则:

$$R_j: \{ \bigcup des(C_i) \rightarrow des(D_j) \mid C_i \cap D_j \neq \emptyset \}$$

2.3 Self 集的构造

2.3.1 特征码的构造

特征码是不定长度的二进制编码。由于所提取的规则都是对安全行为的特征描述,故决策属性默认为“安全”,编码中可以省略。但由于系统中的资源多种多样,对一个资源的访问,不仅有不同的操作,每种操作又有参数,在属性提取阶段应该能针对不同资源的不同操作以及操作的不同参数生成多种特征属性(这里也可以借助现有的专家知识对用户操作类型的分类,用 Rough 集约简算法优化各类已知安全行为规则的条件属性),故特征码的编码中必须有标志不同资源的编码部分。考虑资源的多样性,该部分应留有足够长的编码位数。文中假设取 8 位(见图 1)。另外,为了配合 Self 集的演化,特征码应设置一个可信度编码位,初步设为 8 位,用二进制编码最多可记录 255 条重复记录。特征码的可信度的值是由执行 Rough 集数据约简算法时,同一条记录的重复数目决定的。在 Self 集的演化阶段,可根据特征码的可信度大小判断特征码的优劣。据优胜劣汰的原则保留可信度高的特征码,淘汰可信度低的特征码。

资源 8 位	可信度 8 位	属性 1, 属性 2, ..., 属性 n
--------	---------	-------------------------

图 1 特征码的构成

综上所述,特征码是一种不等长的二进制串,其表现形式为: $[资源,可信度,属性 1,属性 2,\dots,属性 n]$ ^[2]。前面 8 位为资源位,其次 8 位为可信度位;又

假设经约简后的特征属性为 20 个,各属性均取 2 位二进制编码,则每个基因编码总长度为 $8+8+20 \times 2=56$ 位。不同操作规则的特征属性不一定正好相同,故实际构造特征码时应根据约简结果和具体情况适当编码,可得到不等长的特征码。

2.3.2 特征码的编码方式

由上可知,特征码由三部分组成资源位、可信度位和属性位。其中资源位和属性位的编码方式基本相同,都是对属性的二进制编码,下面首先介绍如何对安全规则中的属性进行二进制编码:

经 Rough 集的数据约简获得最简的有效规则集,是对合法用户正常行为特征属性的简化,这些特征属性的取值都是量化的数据。根据专家经验和领域背景知识决定各个条件属性的重要度,用 $[0,1]$ 之间的实数表示,数值越大说明其对决策结果的影响越大。为不失一般性,可按取值的大小将各个属性划分到 4 个区间 $[0,0.25]$ 、 $(0.25,0.5]$ 、 $(0.5,0.75]$ 、 $(0.75,1]$,则每一属性可编码为 2 位二进制位串(00,01,10,11),对应属性所在的 4 个区间。

例如:经过 Rough 集约简后,存在有一条安全规则 $a_1b_2c_3 \rightarrow 安全$;其中条件属性 a_1, b_2, c_3 的属性值分别为 0.3, 0.7, 0.8, 则 a_1 被划分到区间 $(0.25,0.5]$,其二进制编码为 01; b_2 被划分到区间 $(0.5,0.75]$,其二进制编码为 10; c_3 被划分到区间 $(0.75,1]$,其二进制编码为 11。

其次,对可信度位的编码是按照信息可信度的大小进行二进制编码的,其中信息可信度是在对安全数据执行 Rough 集理论的数据约简算法时,记录下来的重复信息的数目,这也是对每条规则可信度的初始化。在 Self 集构造过程中,每匹配到一次该种安全操作类型,可信度加 1,以表示该特征码的有效性,便于进化。

根据专家知识或已知安全操作的安全数据经学习得到初始决策表,通过 Rough 集约简得到有效的最简规则集和信息可信度的值,按照上述方法编码即可形成特征码。

2.3.3 Self 集的构造方法

从海量的通信模式数据、系统安全日志或用户审计记录中抽取最能表达合法正常特征的属性集合,经过粗糙集的数据约简算法,提取出有效的最简规则集和对应规则的可信度,然后进行编码,将最初构造的特征码加入到 Self 集合中,构成最初的 Self 集(见图 2)。

3 Self 集的演化算法设计

经过了 Self 集的构造过程后,现在的 Self 即可以实际应用了。经过一段时间的运行,Self 集就包含了

相当数量正常访问行为的特征码。当然,要包括所有的正常访问序列是不可能的,Self 集还需要不断地演化。

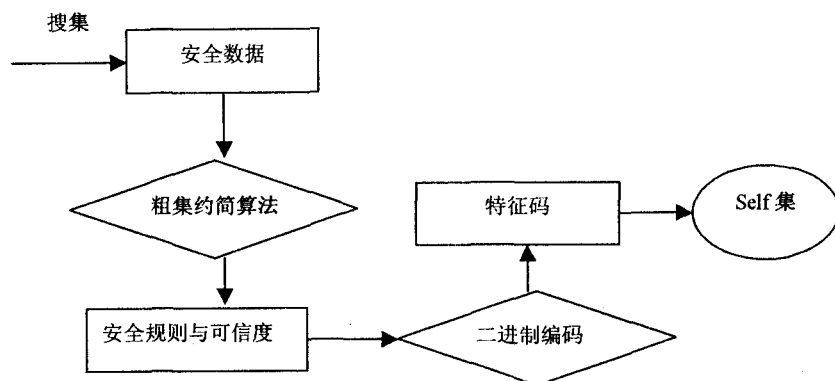


图 2 Self 集的构造方法

下面介绍 Self 集的演化算法。

根据 Rough 集上、下近似集的原理,构造 Self 集的上、下近似 Self 集。

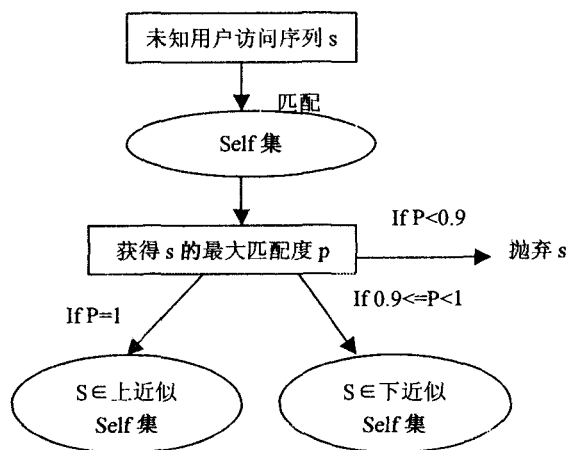
针对一条用户的资源访问序列 s ,不能判断它是合法的还是非法的,这里就要将它与最初的 Self 集作匹配,依据匹配度 p 的值,判断如何处理这个访问序列。

匹配度的计算方法: $P = M/N$

其中: M 是二进制字符串中相同字符的个数(可信度位不参与匹配), N 是二进制字符串的总数 - 8(不包含可信度的 8 位编码)。

3.1 上、下近似 Self 集的构造算法

构造算法参见图 3。



S 的可信度随着匹配次数逐次+1

图 3 上、下近似 Self 集的构造算法

Begin $i = 0$

Next:将 s 与 Self 集的特征码作匹配,搜寻与 s 匹配度最大的规则。计算匹配度 p

If $p = 1$ then

{ If s 存在于上近似 Self 集 then

s 的特征码的可信度 + 1

Else

将 s 添加到上近似 Self 集,并且 s 的特征码的可信度 = 匹配特征码的可信度 + 1

}

Else if $0.9 \leq p < 1$ then

{ If s 存在于下近似 Self 集 then

s 的特征码的可信度 + 1

Else

将 s 添加到下近似 Self 集,并且 s 的特征码的可信度 = 1

}

Else

将 s 删除

End if

$i = i + 1$

Go to Next 如果对 S 分析完毕,则退出

End

经过一段时间的运行,上、下近似 Self 集包含了相当数量的特征码,实验证明,上近似 Self 集中的特征码均属于 Self 集,且其特征码的可信度均高于 Self 集中对应特征码的可信度,可见上近似 Self 集是对 Self 集的进一步优化;而下近似 Self 集中存在的特征码与 Self 集中的特征码有 90% 以上的相似度,可以把下近似 Self 集看作 Self 集的扩展,便于发现新的未知安全特征码。综上可知,上、下近似 Self 集实现了对 Self 集的优化和扩展。

3.2 Self 集的演化算法

(1)利用上、下近似 Self 集的构造算法,构造上、下近似 Self 集。

(2)当上近似 Self 集的特征码数量增长相对缓慢时,更新 Self 集。具体做法:用上近似 Self 集替代原来的 Self 集,上近似 Self 集清空。然后把下近似 Self 集添加到 Self 集中,下近似 Self 集也清空。

(3)转到(1),进入下一个循环的 Self 集演化过程。

该算法实现了 Self 集的动态更新,Self 集演化算法中,通过匹配度控制满足了允许 Self 的合法变形存在,同时,也要识别不合法的改变的要求;符合了在实际的环境中,Self 集与 Nonself 集都不可能是静态的,合法的操作和入侵的方式都是不断变化的实际情况;达到了根据访问的情况自动调整 Self 集的目的。

4 总 结

受人体免疫系统的启发,结合 Rough 集数据约简理论以及 Rough 集的上近似集和下近似集的原理,提出了一种基于 Rough 的 Self 集构造和演化算法。收集

(下转第 67 页)

得到的子类, ZSuper 是直接父类, TSuper 是经推理得到的父类, Extend Brother 是扩展的兄弟类, Extend T-Brother 是扩展后推理得到的兄弟类。 $\Phi_1, \Phi_2, \Phi_3, \Phi_4, \Phi_5, \Phi_6, \Phi_7$ 表示各节点的权值。一般来说 Φ_1 的权值最高, 往后依次降低。 \cup 表示集合并集操作。这样对用户的偏好进行了“放大”, 依照权值的大小, 顺序选择所要搜索的信息, 才能提高查准率和查全率。

最后, 用户偏好库提取了用户的兴趣爱好, 知道它应该选择一些什么样的信息。然后对搜索器从 Internet 抓取回来的信息进行过滤, 步骤如下:

(1) 对于取回来的文本信息或页面, 根据用户个人兴趣剖像中的词条及权重, 对该文本信息或页面进行分值计算^[2]。

$$\text{Score}(\text{URL}_i) = \sum_{i=1}^n \text{Count}(l_i, \text{URL}_i) * W_i$$

$\text{Count}(l_i, \text{URL}_i)$ 为用户个人兴趣剖像中第 i 个词条在 URL_i 所对应的页面中出现的次数, $l_i \in T, T = \{t_1, t_2, \dots, t_m\}$ 为词条集(词典), $W_i \in [0, 1]$ 为 l_i 词条在用户兴趣剖像中的权重, $\sum_{i=1}^n w_i = 1$ 。

(2) 当对搜索器搜索集的 URL 过滤完全后, 按照 $\text{Score}(\text{URL}_i)$ 的分值由高到低依次排序, 这样排序高的最有可能就是用户所需要的。

(上接第 63 页)

尽可能多的安全数据, 经过 Rough 集的数据约简算法, 降低了安全数据的冗余, 减轻了特征码构造的负担, 提取出有效的最简规则集和信息可信度, 将特征属性的取值离散划为 4 个区间, 进行二进制编码构造特征码, 得出最初 Self 集; 同时利用上近似 Self 集和下近似 Self 集的构造实现了 Self 集的演化, 使得 Self 集具有自动演化更新的能力, 为整个免疫系统的构造打下了良好的基础。

当然, 在 Self 集的构造和演化中, 仍有一些关键的问题没有解决, 需要做进一步的工作^[6,7]。

参考文献:

- [1] 李 涛. 计算机免疫学[M]. 北京: 电子工业出版社, 2004: 39-43.
- [2] Hofmeyr S, Forrest S. Architecture For An Artificial Immune System[J]. Evolutionary Computation, 2000, 8(4): 443-473.
- [3] Forrest S, Hofmeyr S A. Immunology as information processing. In: Segel, Cohen eds. Design Principles for the Immune

(3) 最后用户偏好库要进行自我更新, 以便搜索器再搜索时提高查全率。

6 结束语

本体作为一种论点新颖、起点较高, 并且哲学渊源悠久的历史组织体系, 在理论上具备很多优越性和潜在功能。将它应用在智能搜索中, 必然有其独到之处。在智能搜索的用户偏好库模型中, 在本体论的支持下, 采用了用户个人兴趣剖像算法、扩展查询算法, 使得在基于本体的智能搜索中更能体现出用户的兴趣爱好, 能够大大提高检索的速度和查准率。

参考文献:

- [1] 李 景. 本体理论在文献检索中的应用研究[M]. 北京: 北京图书馆出版社, 2005.
- [2] 凌 云, 王 勋, 费玉莲. 智能技术与信息处理[M]. 北京: 科学出版社, 2003.
- [3] 周 宁, 张玉峰, 张李义. 信息可视化与信息检索[M]. 北京: 科学出版社, 2005.
- [4] 徐宝文, 张卫丰. 搜索引擎与信息获取技术[M]. 北京: 清华大学出版社, 2003.
- [5] 秦玄铮. 基于本体的个性化信息检索系统的设计与实现[D]. 北京: 北京邮电大学, 2006.

System and Other Distributed Autonomous Systems. USA: Oxford University Press, 2000: 224-227.

- [4] Jin X Y, Du H F, He W H, et al. Optimizing the weights of neural Networks based on antibody colonel simulated annealing algorithm[C]//The International Symposium Neural Networks (ISNN2004). Heidelberg: Springer - Verlag, 2004: 299-304.
- [5] 梁意文, 李俊涛, 郭学理. 一种基于用户行为的 Self 集构造和演化方法[J]. 计算机应用研究, 2001(9): 7-9.
- [6] Garrett S. From Natural to Artificial Immune Systems[M/OL]. users. aber. ac. uk /smg/Modules/CX211_2001-2002/immunity. ppt, 2004: 110-115.
- [7] Harmer P K, Williams P D, Gunsch G H, et al. An Artificial Immune System Architecture for Computer Security Applications[J]. IEEE Transaction on Evolutionary Computation, 2002(3): 229-335.
- [8] 刘 清. Rough 集及 Rough 推理[M]. 北京: 科学出版社, 2001: 194-201.
- [9] 杨孔雨, 王秀峰. 入侵检测免疫模型中抗体基因库的生成和进化[J]. 计算机应用, 2003(7): 26-28.