人工免疫行为轮廓取证分析方法

杨 珺[1]，曹 阳[1,2]，马秦生[1]，王 敏[3]

(1. 武汉大学电子信息学院 武汉 430079; 2. 武汉大学软件工程国家重点实验室 武汉 430072;
3. 通信指挥学院二系 武汉 430010)

【摘要】针对当前数据挖掘取证分析方法存在的取证分析效率低的问题，提出了采用免疫克隆算法来构建频繁长模式行为轮廓的取证分析方法。该方法以行为数据和频繁项集的候选模式分别作为抗原和抗体，以抗原对抗体的支持度作为亲和度函数，以关键属性作为约束条件，以最小支持度作为筛选条件，通过对抗体进行免疫克隆操作来构建基于频繁长模式的行为轮廓；采用审计数据遍历行为轮廓匹配对比的分析方法来检测异常数据。实验结果表明，与基于Apriori-CGA算法的取证分析方法相比，该方法的行为轮廓建立时间和异常数据检测时间均大幅降低。该方法有助于提高取证分析的效率以及确立重点调查取证的范围。

关 键 词 人工免疫; 行为轮廓; 计算机取证; 计算机安全; 数据挖掘; 电子犯罪对策; 信息分析; 模式匹配
中图分类号 TP309 **文献标识码** A doi:10.3969/j.issn.1001-0548.2010.06.022

Forensic Analysis Method of Behavior Profiling on Artificial Immunity

YANG Jun[1], CAO Yang[1,2], MA Qin-sheng[1], and WANG Min[3]

(1. School of Electronic Information, Wuhan University Wuhan 430079;
2. State Key Laboratory of Software Engineering, Wuhan University Wuhan 430072;
3. Second Department, Commanding Communications Academy Wuhan 430010)

Abstract To improve the efficiency of the forensic analysis method on data mining, this paper proposes a new method for the forensic analysis of the behavior profiling on the longest frequent pattern which is constructed by immune clonal algorithm. Taking the behavior data and the candidate pattern of the frequent item sets as the antigen and the antibody respectively, the support of the antigen to the antibody as the function of affinity, the key attribute as the constraint condition, and the minimal support as the screening condition, the behavior profiling on the longest frequent pattern is built with the help of the immune clonal operation to antibody. The abnormal data are detected by the matching method that the audit data pass through the list items of the behavior profiling. The proposed method and the method on Apriori-CGA are applied in the same problem. The comparison results indicate that the setting up time of behavior profiling and the test time of abnormal data are dramaticly reduced. Therefore, the proposed method has a good ability in the efficiency of forensic analysis and electronic crime investigation.

Key words artificial immunity; behavior profiling; computer forensics; computer security; data mining; electronic crime countermeasures; information analysis; pattern matching

目前，针对行为数据的取证分析主要采用的是异常检测法。该方法通过构造行为轮廓，继而检测审计数据和行为轮廓间的偏差获取证据分布的范围[1-2]。通常，构造行为轮廓需要分析大量的行为数据[3]，因此关联规则挖掘技术被应用于取证分析中。当前，在该领域，实现该技术主要基于Apriori及其改进算法，如文献[4]针对行为数据取证分析提出的Apriori-CGA (criminal behavior profiling generation algorithm)算法，该种算法通过挖掘训练数据的频繁模式构造行为轮廓[4-6]，但由于取证分析过程的时间开销较大，致使取证分析的效率较低，表现为：(1) 算法收敛速度慢，导致行为轮廓建立时间较长；(2) 算法输出无趣模式多，导致行为轮廓规模较大，异常数据检测时间较长。

针对以上问题，本文提出人工免疫行为轮廓的取证分析方法。该方法采用免疫克隆算法挖掘训练数据的频繁模式，以加快算法的收敛速度，减少行

收稿日期：2009 – 06 – 03; 修回日期：2009 – 10 – 14
基金项目：高等学校博士学科点专项科研基金(20040486049)；国家高技术研究发展计划(2002AA1Z1490)
作者简介：杨 珺(1973 –)，女，博士生，主要从事信息安全和软件工程方面的研究.

为轮廓的建立时间；通过设定关键属性来阻止无趣模式的产生，并通过挖掘频繁长模式进一步压缩行为轮廓的规模，以加快异常数据检测的速度。

1 人工免疫行为轮廓取证分析原理

人工免疫行为轮廓取证分析主要分为行为轮廓构造和异常数据检测两个阶段。行为轮廓构造指在安全隔离的环境下，基于正常的训练数据构造用户行为模式的过程，即是以关键属性作为约束条件，以最小支持度作为筛选条件，采用免疫克隆算法从预处理后的训练数据中挖掘频繁长模式的过程，其规模取决于其包含模式的数量；异常数据检测指将行为轮廓和预处理后的审计数据进行比较，查找出两者间相异的数据集，即可疑数据集，从而确定证据可能分布范围的过程。图1给出了人工免疫行为轮廓取证分析的原理框图。

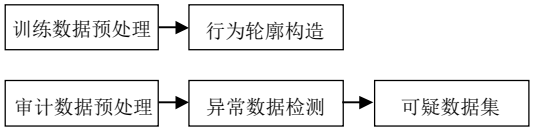

图1 人工免疫行为轮廓取证分析原理

1.1 数据预处理

数据预处理旨在清除源数据中的噪声数据、修复源数据中的不完整数据以及概化源数据中的属性值，并对每个属性值进行编码以满足数据挖掘算法格式的要求。设行为数据的属性集合为 $a=\{a_1,a_2,\cdots,a_n\}$， a 的值域集为 $v=\{v_1,v_2,\cdots,v_n\}$，其中 n 为 a 中属性的数目， v_i 为属性 a_i 的值域，$i=1,2,\cdots,n$。在人工免疫行为轮廓取证分析方法中，经过去噪、修复和概化处理后的行为数据的属性值均采用自然数编码，其中 $a_i=0$，表示 a_i 属性与其他属性无关联。

1.2 行为轮廓构造

行为轮廓采用带关键属性约束的免疫克隆算法来构造。免疫克隆是人工免疫系统理论的重要学说，免疫克隆算法具有收敛速度快、全局和局部搜索能力强的特点[7-9]，因此可以快速地实现频繁模式的挖掘[10]。

关键属性指对描述数据性质起决定性作用的属性。设 $m\leqslant n$， $u=\{u_1,u_2,\cdots,u_m\}$ 是 a 的子集，即 $u\subseteq a$， m 为 u 中属性的数目。若 a 完全函数依赖于 u，则 u 为关键属性集，其任一属性 $u_i\in u$ $(i=1,2,\cdots,m)$ 为关键属性。在行为轮廓构造中，采用关键属性作为挖掘频繁模式的约束条件可以滤除无趣模式[11]，即滤除仅由非关键属性组成的频繁模式，减小行为轮廓的规模。

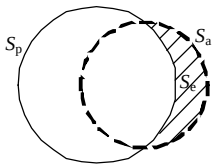

图2 证据分布图

为最大限度地压缩行为轮廓的规模，行为轮廓采用频繁长模式构成。在图2中，设 S_p 为行为轮廓的行为特征集， S_a 为审计数据的行为数据集， s_e 为证据分布的可疑数据集，则有：

$$S_e=S_a-(S_a\bigcap S_p) \tag{1}$$

由式(1)可见，当 S_a 确定时，为缩小 S_e 以降低误检率，应尽可能地增大 S_p，即增大行为轮廓的规模。但是，过大的轮廓规模会增加审计数据和行为轮廓间的比对次数，导致异常数据检测时间开销增大。由于频繁项集的所有非空子集都是频繁的[10,12]，因此可以将频繁模式合并删减，行为轮廓就可以仅由包含了所有频繁模式特征信息的频繁长模式构成，从而达到在缩小行为轮廓规模但不增加误检率的情况下，减少异常数据检测时间的目的。

1.3 异常数据检测

异常数据检测采用审计数据遍历行为轮廓匹配对比分析的方法[10]，检测的速度取决于行为轮廓的规模。设频繁长模式 $Y=\{y_1,y_2,\cdots,y_n\}$，审计数据 $X=\{x_1,x_2,\cdots,x_n\}$，其中 y_i、x_i 分别是 Y、X 的属性项，$i=1,2,\cdots,n$。设距离函数 $L=\sum_{i=1}^{n}w_i\frac{|y_i-x_i|}{v_i}$，当 $x_i=0$ 时，即当 x_i 属性与其他属性无关联时， y_i 取为0； w_i 为属性的权，且有 $\sum_{i=1}^{n}w_i=1$。当距离域值为 $L_{\max}$ 时，则可得 X 为正常数据的充要条件为：

(1) X 中非0属性数≤Y 中非0属性数；

(2) $L\leqslant L_{\max}$。

即在异常数据检测时，不满足上述条件的审计数据将被收集到可疑数据集中。

2 行为轮廓构造算法

行为轮廓采用免疫克隆算法构造，该算法以行为数据和频繁项集的候选模式分别作为抗原和抗体，以抗原对抗体的支持度作为亲和度函数，以关键属性作为约束条件，对抗体进行克隆、变异、重组和选择操作，形成新一代抗体[13]；以最小支持度

min_sup 作为筛选条件，从抗体中输出频繁长模式。

2.1 亲和度函数

亲和度指抗原和抗体匹配的程度，亲和度的大小由亲和度函数度量。在免疫克隆算法中，抗体的支持度指抗原中包含抗体的数量与抗原总数之比，说明了抗体在抗原中的代表性，大小反映了抗体成为频繁模式可能性的大小[10]。因此，在行为轮廓构造中，可选用抗体的支持度作为亲和度函数。设 $\sup(A_i)$ 为抗体 A_i 的支持度，$f(A_i)$ 为抗体 A_i 的亲和度函数，则 $f(A_i)=\sup(A_i)$。

2.2 克隆操作

设 k 为克隆代数，N 为抗体种群规模，A_i 为抗体，$A(k)$ 为 k 代抗体种群，有 $A(k)=[A_1(k),A_2(k),\cdots,A_N(k)]$，则克隆操作可定义为 $T_c(A(k))=[T_c(A_1(k)),T_c(A_2(k)),\cdots,T_c(A_N(k))]$。其中，$T_c(A_i)=A_i\times\mathbf{I}_i$ $(i=1,2,\cdots,N)$，$\mathbf{I}_i$ 为 q_i 维全1行向量。q_i 的大小取决于亲和度函数和克隆规模，设克隆规模为 N_c，且有 $N_c>N$，则：

$$q_i=\mathrm{Int}\left[N_c\times\frac{f(A_i)}{\sum_{j=1}^{N}f(A_j)}\right]\quad i=1,2,\cdots,N \tag{2}$$

克隆后的抗体种群为 $A'(k)=[A(k),A'_1(k),A'_2(k),\cdots,A'_N(k)]$，其中 $A'_i(k)=[A'_{i1}(k),A'_{i2}(k),\cdots,A'_{i(q_i-1)}(k)]$，$A'_{ij}=A_i$，$j=1,2,\cdots,q_i-1$。

2.3 克隆变异

依据概率 P_m^i 对克隆后的抗体种群 $A'(k)$ 采用随机取值的方式进行变异操作。为了保留抗体原始种群信息，变异算子不作用于 $A(k)\in A'(k)$。在行为轮廓构造中，重复模式对找寻频繁模式无实际意义，所以 $A'_i(k)$ 中的抗体将全部参与变异操作，即变异操作为 $P_m(A'_{ij}(k)\to A''_{ij}(k))=p_m^i=1$，$i=1,2,\cdots,N$；$j=1,2,\cdots,q_i-1$。变异后的抗体种群为 $A''(k)=[A(k),A''_1(k),A''_2(k),\cdots,A''_N(k)]$。

2.4 克隆重组

依据概率 p_c 对变异后的抗体种群 $A''(k)$ 采用单点交叉的方式进行重组操作。为了保留抗体原始种群信息，重组算子不作用于 $A(k)\in A''(k)$，即重组操作为 $P_c((A''_{ij}(k),A''_{st}(k))\to(A'''_{ij}(k),A'''_{st}(k)))=p_c$，$i,s=1,2,\cdots,N$；$j=1,2,\cdots,q_i-1$；$t=1,2,\cdots,q_s-1$。重组后的抗体种群为 $A'''(k)=[A(k),A'''_1(k),A'''_2(k),\cdots,A'''_N(k)]$。

2.5 克隆选择

设新抗体 $B_i(k)=\max\{f(A'''_{ij}(k))\}$，$i=1,2,\cdots,N$；$j=1,2,\cdots,q_i-1$。则当 $f(B_i(k))\geqslant f(A_i(k))$ 时，$B_i(k)$ 取代 $A_i(k)\in A(k)$。选择后的抗体种群为 $A(k+1)=[A_1(k+1),A_2(k+1),\cdots,A_N(k+1)]$。在 $A(k+1)$ 中，若有 $A_i(k+1)=A_j(k+1)$，则删除 $A_j(k+1)$，并在 $A(k+1)$ 中随机生成一个新的无重复的抗体。将 $A(k+1)$ 中满足关键属性约束、满足 $\sup(A_i(k+1))\geqslant\min\_\sup$、且满足频繁长模式条件的当前抗体添加到行为轮廓中。

2.6 算法流程

行为轮廓构造的算法流程如图3所示。初始抗体种群 $A(0)$ 是从抗原集中随机生成的，随机抽取抗原集中的一个抗原，并随机置0该抗原的若干属性位便可得到了一个初始抗体。若 $A(0)$ 中无相同个体，则添加该抗体到 $A(0)$ 中，重复以上操作，直到种群规模满足要求。算法终止操作的条件是克隆代数达到初始设定值 k 或连续若干代操作无新的频繁长模式输出。

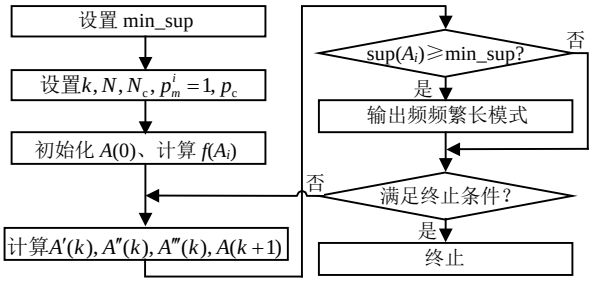

图3　行为轮廓构造算法流程图

3 实验结果及分析

实验数据选自文献[14]的数据文件lbl-conn-7.tar.Z，该组数据共有9个属性，反映了劳伦斯-伯克利实验室(Lawrence-Berkeley laboratory)30天的TCP连接情况。选取protocol和remote host作为关键属性项，以使行为轮廓能够反映用户在网络连接中的行为。顺序截取用户1的10 000条正常连接记录作为训练数据；随机截取用户1剩余记录中的5 000条记录作为审计数据。设 k=1 000，N=300，N_c=900，p_m^i=1，p_c=0.5，min_sup 分别为10%、5%、1%、0.5%和0.1%、0.05%和0.01%。

表1列出了在上述条件下Apriori-CGA算法(方法1)和无约束的免疫克隆算法(方法2)挖掘出的频繁模式数目，以及带约束的免疫克隆算法(方法3)挖掘出的频繁长模式数目；图4描述了在上述条件下方

法1和方法3的运行时间随最小支持度在常用对数坐标下的变化趋势。分析表和图中的信息(10次操作的平均值)可以得到以下结果：

(1) min_sup 应合理选取。

min_sup 选取过大，会抑止有趣模式的产生，导致行为轮廓模型不完整，证据的分布范围过宽，取证分析误检率增大；而 min_sup 选取过小，算法提取的模式数量会急剧增加，如 min_sup = 0.01% 时，相当比例的输出模式是偶然行为产生的，它们并不具有行为特征性，导致异常数据检测时间增长，取证分析效率降低。

(2) 免疫克隆算法较Apriori-CGA算法提取的频繁模式数目略少，但其运行效率却远高于后者。

Apriori-CGA算法采用逐层搜索迭代的方法寻找频繁项集，即用递推的方法由频繁 $i-1$ 项集产生 i 项集[10]，尽管能够生成所有的频繁模式，然而其运算量却较大；而免疫克隆算法采用多点和随机的群体搜索方法来寻找频繁项集，能够快速地收敛于全局最优解[9-10]。由图4可知，免疫克隆算法的收敛时间远低于Apriori-CGA算法的收敛时间，即采用免疫克隆算法可以极大地减小行为轮廓的建立时间。

(3) 采用带关键属性约束的频繁长模式挖掘算法，可以极大地滤除无趣模式和冗余模式，减小行为轮廓的规模，提高异常数据检测的效率。

表1 各种方法挖掘出的频繁(长)模式数目

最小支持度/(%)	方法1	方法2	方法3
10	37	37	1
5	43	43	1
1	163	162.4	8
0.5	397	392.6	24.7
0.1	2 245	2 179.9	169.8
0.05	3 778	3 634.4	286.5
0.01	1 6456	1 5603.2	1 512.4

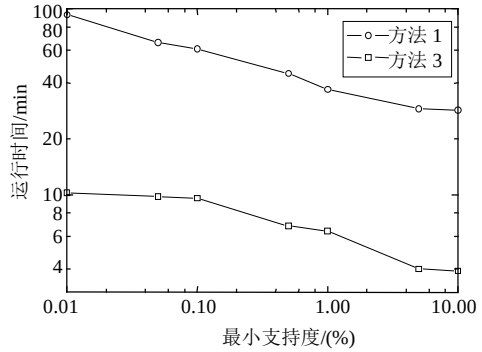

图4 算法运行时间和最小支持度的关系

假设每个属性均为布尔型，则由 n 个属性决定的所有可能的模式数量为：

$$S(n)=\sum_{i=1}^{n}C_n^i=C_n^1+C_n^2+\cdots+C_n^n \tag{3}$$

由式(3)可以算出 $S(5)=31$、$S(9)=511$、$S(13)=8\,191$。可见，随着 n 的增加，模式数量将急剧增长。若不对属性项施加约束，将会挖掘出许多无趣模式；若不对频繁模式进行删减，将会挖掘出许多冗余模式，从而延长确定证据的时间。

表2列出了在 min_sup $=0.1\%$、$w_1=w_2=\cdots=w_n=\dfrac{1}{n}$、$L=L_{\max}=0$ 时，分别采用基于Apriori-CGA算法的取证分析方法和人工免疫行为轮廓取证分析方法(10次操作的平均值)进行异常数据检测提取的特征参数。

表2 异常数据检测结果

特征参数	基于Apriori-CGA算法的取证分析方法	人工免疫行为轮廓取证分析方法
可疑数据/条	1 526	1 544.7
误检率/(%)	11.45	11.92
检测时间/s	6.11	0.94

分析表中的信息可以得到以下结果：Apriori-CGA算法理论上能够完全提取训练数据的频繁模式，然而由于训练数据集规模有限，由训练数据构造的行为轮廓并非是完备的，因此，审计数据中就可能含有行为轮廓未能覆盖的正常模式，从而导致误检率不为0。从表中可以看出，与基于Apriori-CGA算法的取证分析方法相比，人工免疫行为轮廓取证分析方法的可疑数据规模稍大、误检率也稍高，然而检测时间却有了大幅度的降低。因此人工免疫行为轮廓取证分析方法能够在很好体现基于Apriori-CGA算法的取证分析方法主要功能特性的基础上，大幅度地提高取证分析的效率。

4 结 论

取证分析是从海量的数据中获取计算机犯罪证据的技术，是计算机取证技术中最重要的环节。本文通过分析取证数据的特点，提出了人工免疫行为轮廓取证分析方法。该方法充分利用了免疫克隆算法收敛速度快、全局及局部搜索能力强、关键属性对取证数据性质约束能力较强，以及频繁长模式对数据检测空间压缩能力较强的特点，实现了行为轮廓的准确、快速、高效构建，以及异常数据的快速检测。仿真实验表明，该方法能有效地提高取证分析的效率和确立重点调查取证的范围。

(下转第919页)

labeled data by local Fisher discriminant analysis[J]. Journal of Machine Learning Research, 2007, 8: 1027-1061.
[8] ZELNIK M L, PERONA P. Self-tuning spectral clustering [C]//Advances in Neural Information Processing Systems. Cambridge: MIT Press, 2005: 1601-1608.
[9] YAN S, XU D, ZHANG B, et al. Graph embedding and extensions: a general framework for dimensionality reduction[J]. IEEE Trans Pattern Analysis and Machine Intelligence, 2007 , 29(1): 40-51.
[10] GEORGHIADES A S, BELHUMEUR P N, KRIEGMAN D J. From few to many: Illumination cone models for face recognition under variable lighting and pose[J]. IEEE Trans Pattern Analysis and Machine Intelligence, 2001, 3(6): 643-660.
[11] GRAHAM D B, ALLINSON N M. Face recognition: from theory to applications[M]. [S.l.]: Springer, 1998: 446-456.

编 辑 蒋 晓

(上接第914页)

参 考 文 献

[1] PEISERT S, BISHOP M, KARIN S, et al. Analysis of computer intrusions using sequences of function calls[J]. IEEE Trans on Dependable and Secure Computing, 2007, 4(2): 137-150.
[2] MA Xin-xin, ZHAO Yang, QIN Zhi-guang. Improving resilience against DDoS attack in unstructured P2P networks[J]. Journal of Electronic Science and Technology of China, 2007, 5(1): 18-22.
[3] HERRERIAS J, GOMEZ R. A log correlation model to support the evidence search process in a forensic investigation[C]// Proceedings of the Second International Workshop on Systematic Approaches to Digital Forensic Engineering. New York: IEEE Computer Society Press, 2007: 31-42.
[4] 孙 波. 计算机取证方法关键问题研究[D]. 北京: 中国科学院软件研究所, 2004.
SUN Bo. Research on key aspects of computer forensic methods[D]. Beijing: Institute of Software of Chinese Academy of Sciences, 2004.
[5] ABRAHAM T, VEL O. Investigative profiling with computer forensic log data and association rules[C]// Proceedings of the 2002 IEEE International Conference on Data Mining. New York: IEEE Computer Society Press, 2002: 11-18.
[6] ABRAHAM T. Event sequence mining to develop profiles for computer forensic investigation purposes[C]//Proceedings of the 2006 Australasian workshops on Grid computing and e-research. Darlinghurst, Australia: Australian Computer Society, 2006: 145-153.
[7] CASTRO L N, ZUBEN F J. The clonal selection algorithm with engineering applications[C]//Proceedings of GECCO'00, Workshop on Artificial Immune Systems and Their Applications. New York: ACM Press, 2000: 36-37.
[8] TIMMIS J, HONE A, STIBOR T, et al. Theoretical advances in artificial immune systems[J]. Theoretical Computer Science, 2008, 403(1): 11-32.
[9] KHILWANI N, PRAKASH A, SHANKAR R, et al. Fast clonal algorithm[J]. Engineering Applications of Artificial Intelligence, 2008, 21(1): 106-128.
[10] 焦李成, 刘 芳, 刘 静, 等. 智能数据挖掘与知识发现[M]. 西安: 西安电子科技大学出版社, 2006.
JIAO Li-cheng, LIU Fang, LIU Jing, et al. Intelligent data mining and knowledge discovery[M]. Xi'an: Xidian University Press, 2006.
[11] 金可仲. 基于关键属性约束的关联规则挖掘在日志分析中的应用[J]. 温州大学学报, 2008, 29(1): 56-60.
JIN Ke-zhong. Application on log analysis using association rule mining algorithm with key item constraint[J]. Journal of Wenzhou University, 2008, 29(1): 56-60.
[12] AGRAWAL R, SRIKANT R. Fast algorithm for mining association rules[C]//Proceedings of the 20th Very Large Data Bases International Conference. San Francisco: Morgan Kaufmann Publishers, 1994: 487-499.
[13] 杨 珺, 王 敏, 陈 晨, 等. 带约束的免疫克隆取证分析方法[J]. 计算机工程, 2010, 36(14): 158-160.
YANG Jun, WANG Min, CHEN Chen, et al. Immune clone forensic analysis method with constraint[J]. Computer Engineering, 2010, 36(14): 158-160.
[14] PAXSON V. Traces available in the internet traffic archive[EB/OL]. (2008-04-09) [2008-07-30]. http://ita.ee.lbl.gov/htm /traces.html.

编 辑 张 俊