遗传算法和关联规则算法用于入侵检测系统的研究*)

黄 勤 贺向前 刘益良 姚雪梅

(重庆大学自动化学院 重庆 400000)

摘 要 基于异常的入侵检测系统常采用关联规则挖掘算法,关联规则算法的最小支持度和最小置信度设置不仅要影响入侵检测系统的检出率和虚警率,还要影响入侵检测系统的负荷。本文提出遗传算法搜寻关联规则算法最小支持度和最小置信度最优的设置范围,为实时的入侵检测系统的关联规则挖掘算法提供参数参考,改善入侵检测系统的实时性,提高检出率,降低虚警率。

关键词 关联规则,遗传算法,入侵检测

1 引言

基于异常的入侵检测系统常采用关联规则挖掘算法。关联规则算法的最小支持度和最小置信度设置不仅要影响入侵检测系统的检出率和虚警率,还要影响入侵检测系统的负荷。

图1是一个典型的入侵检测系统模型,其中规则的生成是关键。关联规则算法广泛应用于基于异常检测(anomaly detection)的入侵检测系统规则的生成,其核心技术是如何挖掘频繁项集(Frequent Itemset)中的规则。

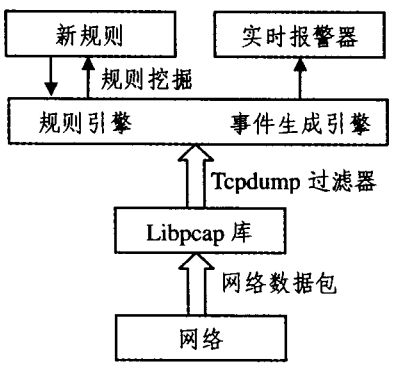

图1 入侵检测系统模型

入侵检测的实现方法很多,文[1]提出了数据挖掘的方法来建立入侵检测系统,介绍了关联规则算法和频繁序列算法(Frequent Episodes);文[2]讨论了基于关联规则的入侵检测系统中如何改进算法,提高挖掘效率;文[4]引入了分布级别的概念,提出了一种层次化协作模型(HCM)。实际的入侵检测系统属于多种算法的综合应用,其中关联规则算法对于基于异常的入侵检测系统非常适用。文[3]主要关注关联规则算法的实现细节和数据结构,对关联规则挖掘中最小支持度和最小置信度的阈值设置问题讨论很少。事实上,最小支持度和置信度的设置是影响关联规则算法效率的关键因素,从而影响入侵检测系统的检出率和虚警率以及入侵检测系统的运算量和系统负荷。本文利用遗传算法(Genetic Algorithm)的优化搜索功能,搜寻关联规则算法的最优参数,即最小支持度和最小置信度的合理设置范围,改善了入侵检测系统的实时性,提高了检出率,并在降低虚警率方面作了进一步研究。

2 入侵检测系统中的关联规则算法

规则的生成是入侵检测系统的核心。目前,规则的生成主要采用关联规则(Association Rules)算法实现,关联规则算法的主要思想可有两步:

(1)产生频繁项集。

(2)由频繁项集生成关联规则。

其中,第二步生成关联规则实现相对容易;第一步产生频繁项集实现较复杂,其算法思想可如图2流程所示。

图2中 L_k 代表算法中的频繁项集,C_k 代表算法中的候选集,Sup(I)为候选集第I项的支持度,Min_Sup 为算法设置的最小支持度。

该算法的负荷很大。如果单纯提高最小支持度的阈值 Min_Sup,候选集中满足条件"Sup(I)>=Min_Sup"项会减少,因此由候选集产生频繁集的规模就会减小,即算法中由"$L_k = L_k + C_{k+1}$ 中的第I项"产生的频繁集 L_k 的数量会减少,这样也导致了 L_k 自连接产生候选集 C_{k+1} 的量减少,从而数据库扫描次数减少,计算的收敛速度加快,由此,入侵检测系统的实时性可以得到保证,检出率大大增加,但虚警率也将同时提高。为解决这一矛盾,只能在最小

*)本文研究得到重庆市自然科学基金项目(CSTC,2004BB2181)资助。黄 勤 副教授,硕士生导师,研究方向为信息安全领域和计算机硬件技术;贺向前 讲师,在读研究生,研究方向为医学信息处理;刘益良 教授,研究方向为信息安全领域;姚雪梅 讲师,在读研究生。

支持度 Min_Sup 和最小置信度 Min_Conf 的设置与实时性、检出率和虚警率之间寻求理想的值。利用遗传算法则可实现对其寻优的目的。

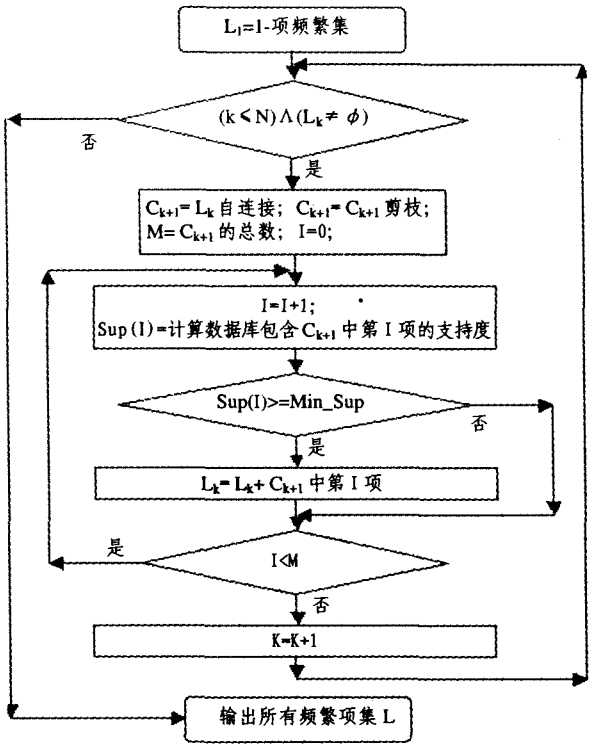

图 2　Apriori 算法

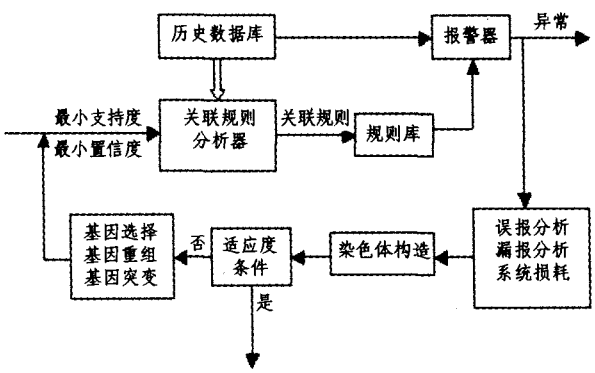

图 3　遗传算法寻优关联规则算法的参数

3　遗传算法寻优关联规则算法参数

利用遗传算法寻优关联规则算法的参数的流程图如图 3 所示。此方法的作用在于事后分析，引入遗传算法可以获得最小支持度和最小置信度阈值的合理设置范围，为实时入侵检测系统的关联规则算法提供参数设置参考。事后分析是单独进行的，只需要少量硬件进行分析演算，普通 PC 机即可，不会影响入侵检测系统的实时性。

3.1　染色体的产生和编码

用最小支持度和置信度等作为染色体的基因片段，采用二进制编码组成一个染色体，初始随机产生 n(n=50)个染色体的物种经过反复的基因选择、重组和突变，可以产生优良的染色体。如：

最小支持度(0<=Min_Sup<=1)扩大 10^5 后采用 16 位二进制编码为：

0001 1111 0100 0000(80%)

最小置信度(0<=Min_Conf<=1)扩大 10^5 后采用 16 位二进制编码为：

0001 0011 1000 1000(50%)

则种群 p 的一个染色体 32 位编码为：

0001 1111 0100 0000 0001 0011 1000 1000

3.2　适应度 f 的计算及基因优化

对于适应度的计算，需要构造目标函数。最小支持度和最小置信度的设置决定入侵检测系统的误报率(R_misInfo)、漏报率(R_falseReport)和系统损耗(T)三因素，因此可以把这三种因素的复合作用作为目标函数，即优化的对象。所以适应度函数可采用如下的方式构造：

$$f=\text{fitness}(R\_misInfo, R\_falseReport, T)$$
$$=\omega_1 * F_1(R\_misInfo)+\omega_2 * F_2(R\_falseReport)+\omega_3 * F_3(T)$$

其中 $\omega_1\in[0,1]$为误报率 F_1(R_misInfo) 的权重因子，$\omega_2\in[0,1]$为漏报率 F_2(R_falseReport)的权重因子，$\omega_3\in[0,1]$为系统损耗 F_3(T)的权重因子，且有 $\omega_1+\omega_2+\omega_3=1$。

3.3　基因选择(Selection)

种群中适应能力强的个体获得更多的选择机会，因而产生的下一代可以获得更高的适应度，使系统向最优的方向发展。

在此采用轮盘选择法，其方法是计算种群中所有染色体的适应度值的总和[S]，然后在[0,S]的搜索空间中随机产生一个 R，选择适应度值大于 R 的染色体[6]。

3.4　基因重组(Crossover or Recombination)

基因重组就是通过两个父染色体杂交产生两个子染色体。在此基因重组的概率设为 0.9，少部分染色体直接复制父染色体，采用单点杂交，其过程如下：

确定染色体编码串中的某一点 P，复制第一个染色体从开始到 P 点的信息和第二个染色体从 P 点到末尾的信息组合成一个新的染色体，再复制第一个染色体从 P 点到末尾的信息和第二个染色体从开始到 P 点的信息组合成另一个新的染色体，如

父 A：

0001 1111 0100 0000 0001 0011 1000 1000

父 B：

0001 0000 0101 0100 0001 0011 1000 1000

设 P 点在 8 位，则单点杂交后的新染色体为：

子 A：

0001 1111 0101 0100 0001 0011 1000 1000

子 B：

0001 0000 0100 0000 0001 0011 1000 1000

3.5 基因突变(Mutation)

基因突变就是随机将染色体的二进制编码串中的任意位二进制数翻转形成新的染色体，可防止局部收敛。随机改变基因的概率一般不适宜设置过高，通常为0.05[7]，其突变过程如下：

原染色体：

1000 0011 0100 0000 0001 0011 1000 1000

基因突变：

1000 0011 0000 0100 0001 0011 1000 1000

3.6 迭代计算，优化目标

通过择优选择、基因重组和突变后产生了新的物种，新的物种是向目标优化的方向进行的，这样的寻优迭代计算反复进行，直到选出个体的适应度达到允许的优化值，或其他收敛条件[7]。

4 实验仿真与数值分析

实验采用了KDD Cup 1999提供的数据[5]，导入数据到SQL Server数据库。针对SYN Flood攻击，对数据进行预处理。实验环境为：P4 2.0GHZ处理器，256MB内存，Win2000操作系统，SQL Server2000数据库，Java开发工具，最小置信度固定设置为50%。表1为实验结果。

对表1中的数据进行分析，发现如图4所示的变化趋势，即当遗传算法迭代运行到第50代以后，最小支持度在50%左右波动，检出率在74%左右，虚警率为1.5%左右。

表1 支持度VS检出率、虚警率和系统损耗

遗传算法代数	支持度	检出率	虚警率	运行时间(秒)	内存消耗MB
第5代	80%	89%	4%	0.8	59.7
第10代	60%	79.9%	3%	1.2	61.2
第15代	27%	29.50%	0.67%	26.3	118.4
第20代	59%	78.8%	2.7%	1.4	62
第25代	34%	31.20%	0.89%	14.1	108
第30代	37%	59%	1.02%	8.6	80.3
第35代	55%	77%	2%	1.5	63.9
第40代	52%	76%	1.82%	1.7	67.8
第45代	49%	74.30%	1.30%	2.9	70.3
第50代	50%	74%	1.5%	2.1	70.1
第55代	51%	75.20%	1.70%	1.9	70

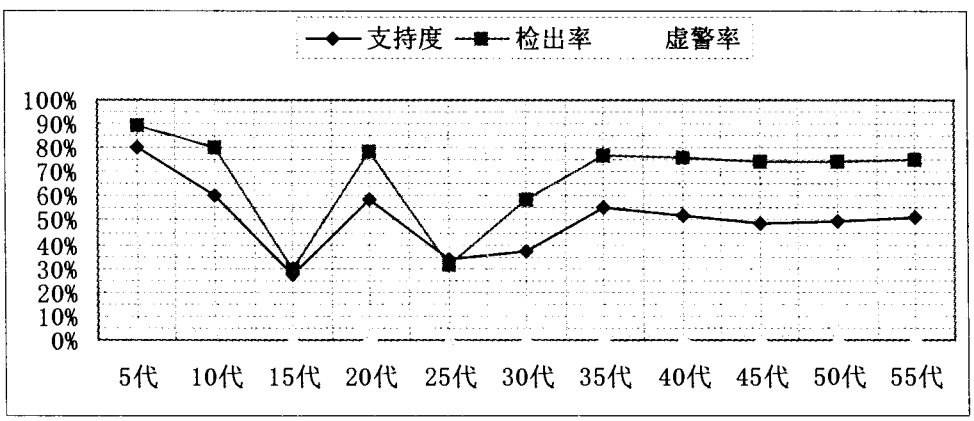

图4 遗传算法代数VS支持度、检出率和虚警率

同时分析表1还发现，遗传算法运行到50代左右时，其运行时间为2秒左右。如果对实验数据根据最小支持度排序，支持度较大时，系统运行时间较短，随着支持度的减小，系统运行时间急剧增加，近似指数分布 $t(s)=\lambda e^{-\lambda s}$，趋势变化如图5所示。

结论 通过遗传算法搜索关联规则挖掘算法参数，可以为入侵检测系统的关联规则挖掘算法提供参数参考，实验结果表明，当最小支持度为50%时，能够有效平衡入侵检测系统实时性、检出率和虚警率之间的矛盾。实验中考虑到运算的负荷，在遗传算法的每次迭代运算中，采用了统一的最小置信度设置(50%)，如果将对每次迭代中采用相同的最小支持度和非50%最小置信度进行模拟，将得到每次迭代采用相异的最小支持度和最小置信度的实验结果。另外，入侵检测系统关联规则通常是多层次的，如果最小支持度阈值过高，则容易忽略低层次项中有意义的关联关系，如果设置过低，会产生许多无意义的关联关系。对于不同的层次采用不同的最小支持度阈值设置问题，其方法还在进一步研究解决之中。

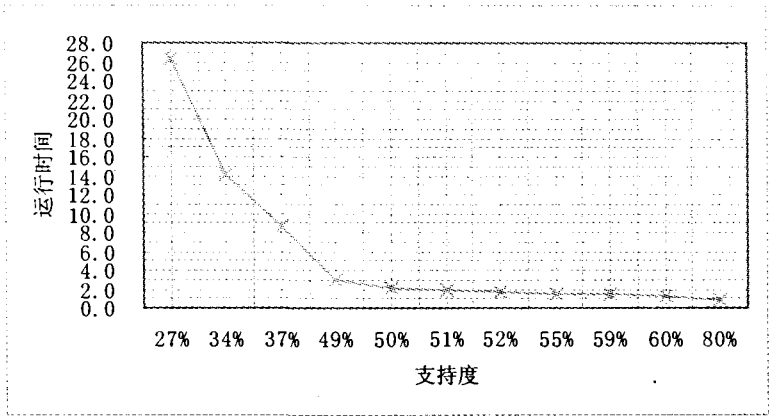

图5 系统运行时间 VS 支持度

参 考 文 献

1 Lee W, Stolfo S J. Data Mining Approaches for Intrusion Detection. http://www1. cs. columbia. edu /~ sal/hpapers/USENIX/usenix. html

2 朱秋萍,毛平平,罗俊. 基于关联规则的入侵检测系统. 计算机工程与应用,2004,26:160~173

3 Agrawal R, Srikant R. Fast algorithms for mining association rules. http://citeseer. csail. mit. edu/cache/papers/cs/1451/http: zSzzSzwww. almaden. ibm. comzSzcszSzpeoplezSzragrawalzSzpaperszSzvldb94_rj. pdf/agrawal94fast. pdf

4 连一峰,戴英侠,胡 艳,许一凡. 分布式入侵检测模型研究. 计算机研究与发展,2003,40(8)

5 KDD Cup 1999Data. http://kdd. ics. uci. edu/databases/kddcup99/kddcup. data. gz

6 Mitchell M. An introduction to genetic algorithms. Cambridge, MA: The MIT Press, 1996

7 Rudolph G. Convergence analysis of canonical genetic algorithms. IEEE Transactions on Neural Networks, 1994, 5(1): 86~101

(上接第 49 页)

恶意用户发送大量复杂的 MIME 结构的消息到代理服务器,恶意代理服务器发送大量复杂的 MIME 结构的包含源 IP 和邻近代理服务器的 Via 头消息来使代理服务器不能提供有效服务及总体性能下降。基于此代理服务器必须支持对用户认证免受重放攻击、对服务器认证免受重放攻击,支持用户认证和数据完整性保护、服务器认证和数据完整性保护,防止重放攻击和拒绝服务攻击。

结束语 随着 SIP 技术的不断成熟和应用的不断发展,SIP 的安全性将会得到越来越深入的研究。本文从代理服务器的角度分析了 SIP 体系结构中 UAC 与代理服务器之间、代理服务器自身以及代理服务器与代理服务器之间现有的安全策略,并提出了用事务的概念及协商机制来实现 SIP 安全性的策略。这种安全策略将会占用代理服务器的部分资源并对性能有一定影响,因此还需要根据实际情况采取相应程度的安全策略,并在实施的过程中考虑人为因素的影响,如管理员及终端用户的干涉。

参 考 文 献

1 Rosenberg J, Schulzrinne H, Camarillo G. SIP: Session Initiation Protocol IETF, RFC 3261, June 2002

2 Burger E, Van Dyke J, Spitzer A. Basic Network Media Services with SIP IETF, RFC4240, December 2005

3 Ono K, Tachimoto S, Corporation NTT. Requirements for End-to-Middle Security for the Session Initiation Protocol (SIP) RFC 4189, Octorber 2005

4 Dierks T, Independent E, Rescorla RTFM, Inc. The Transport Layer Security (TLS) Protocol(Version 1.1) RFC 4346, April 2006

5 Blaze AT M, Labs T A. Keromytis Columbia University M. Richardson Sandelman Software Works L. Sanchez Xapiens Corporation IP Security Policy (IPSP) Requirements RFC 3586, August 2003

6 Arkko J, Torvinen V, Camarillo G, Ericsson, Niemi A, Haukka T, Nokia. Security Mechanism Agreement for the Session Initiation Protocol (SIP). RFC3329, January 2003

7 Rosenberg J, Systems C, Schulzrinne H, Columbia University. Guidelines for Authors of Extensions to the Session Initiation Protocol (SIP), RFC4485, May 2006

8 Camarillo G 著. 白建军,彭晖,田敏译. SIP 揭密[M]. 北京:人民邮电出版社,2003

9 Poikselka M, Mayer G, Khartabil, Niemi A 著. 赵鹏,周胜,望玉梅译. IMS:移动领域的 IP 多媒体概念和服务[M]. 北京:机械工业出版社,2005

10 张智江,张云勇,刘韵洁. SIP 协议及其应用[M]. 北京:电子工业出版社,2005

11 卿斯汉. 安全协议[M]. 北京:清华大学出版社,2005

12 Pfleeger C P 著. 李毅超,等译. 信息安全原理与应用[M]. 北京:电子工业出版社,2004