遗传算法在入侵检测中的应用*

黄 羽** 黄迪明 何险峰 武 明

(电子科技大学计算机科学与工程学院 成都 610054)

【摘要】介绍了基于模型推理和基于模型两种入侵检测系统，提出了一种新的基于智能体技术的入侵检测系统体系结构，解决了传统集中式入侵检测系统的弊病，将任务处理和数据分布到网络各个结点上，充分利用网络资源协同完成入侵检测任务；介绍了遗传算法在该系统中的应用，因系统安全的先验知识体现在对原始数据中有价值特征属性变量集的选择上，故利用遗传算法对特征属性变量子集的选择进行优化，找到相对最优的由特征向量表示的特征属性变量集，以降低入侵检测系统的负荷。

关 键 词 遗传算法; 适应度; 入侵检测; 数据挖掘; 特征向量; 特征子集

中图分类号 TP316 文献标识码 A

Research of Genetic Algorithm Applied to IDS

Huang Yu Huang Diming He Xianfeng Wu Ming

(School of Computer Science and Engineering, UEST of China Chengdu 610054)

Abstract This paper introduces the model discursion-based intrusion detection system and the model-based intrusion detection system and presents a new kind of IDS based on agent, by which IDS distributes data and task to the nodes in the networks. Thus IDS can make best use of compute capability and resources of the networks, which covers the shortage of conventional centralized intrusion detection approach. Importantly, the genetic algorithm applied to the IDS is introduced in detail. In allusion to the apriori knowledge of system security always embodying as the selection of the useful subset of attributes in original data, this IDS uses the genetic algorithms to optimize the feature subset selection and to find the relative optimal subset of attributes expressed by feature vector. The IDS uses data mining algorithms to abstract key features of system runtime status from security audit data, and it uses genetic algorithm to select the feature subset to reduce the amount of data that must be obtained from running processes and classified.

Key words genetic algorithms; fitness; intrusion detection; data mining; feature vector; feature subset

1 入侵检测系统

1.1 基于模型推理的入侵检测

传统的入侵检测系统主要使用指纹识别方法去寻找恶意用户。指纹识别需要对一个电脑系统的所有攻击进行分类，对所有类型找出其独特的特性，并且作一个总和归纳。将各个生成的指纹添加入检测系统的

2003年4月15日收稿

* 总装备部预研基金资助项目

** 男 26岁 硕士 主要从事网络安全及网络多媒体技术方面的研究

攻击数据库，可以用指纹对后继的用户连接进行比较，从而辨认出正常连接或恶意连接。一般来说，指纹总和由系统设计者通过人工分析来完成，而指纹数据的升级则必须通过人工安装到各个使用该系统的机器上。但是，指纹识别方法也存在下述问题：在第一次生成某种指纹时，系统必然要承受该种类型的攻击；每一种不同类型的攻击对应于一种指纹；指纹数目越多，用户检测的系统资源越多，系统性能下降；在一种新的攻击被发现，到对应的新指纹生成的期间，系统将暴露在该种新攻击下；在特定情况下，由于指纹的数目过大，基于指纹的系统不能将所需要的资源用于攻击检测，从而导致无法检测到某些攻击。因而，出现了新的基于模型的入侵检测系统。该系统不再用指纹识别对攻击分类，而是从理论行为模型中得到的用户特征与系统的所有用户进行比较，从而找出攻击者。用户行为一般定义为描述客户端与服务器端之间连接的客观特征集，使用综合行为模型在理论上比指纹识别系统更加准确、有效和易于维护。

1.2 基于模型的入侵检测系统

由于恶意行为本身与正常行为不同，基于模型的入侵检测系统采用的检测方法对某种攻击不需要预先知道，排除了暴露在新攻击下的可能。另外，该系统对每个用户分配一定数量的系统资源，大幅度地降低了耗尽可用资源的可能性。而且，不需要对攻击类型的数据进行升级，因为对该系统的攻击的特性在系统的整个生存期中是不变的。

在以往的系统中，模型生成有两种选择：基于正常用户和基于恶意用户。基于正常用户的模型称为异常检测模型，根据正常用户行为建立模型，对于不匹配该模型的计算机行为活动认为是恶意的；基于恶意用户的模型称为误用检测模型，该模型寻找攻击行为的模式，对于匹配该模式的行为认为是恶意攻击。此外，可以利用遗传算法生成最好的模型，而不直接采用上述两种模型。

入侵检测系统必须能够检测出恶意用户连接，为了达到这个目的，必须有一个该系统用户行为的概括模型。而生成用户模型最有效的方法是应用数据分析算法与训练数据，生成对应于不同用户的理论模型，这里的训练数据代表了真实情况的数据。以往研究表明，数据分析算法包括数据挖掘技术、分散马尔可夫转换和遗传算法等。

1.3 一种新的基于智能体技术的入侵检测系统

在对当前入侵检测系统的体系结构进行分析的基础上，本文提出了一种新的基于智能体(Agent)的入侵检测体系结构，并设计了基于该体系结构的原型系统[1]。

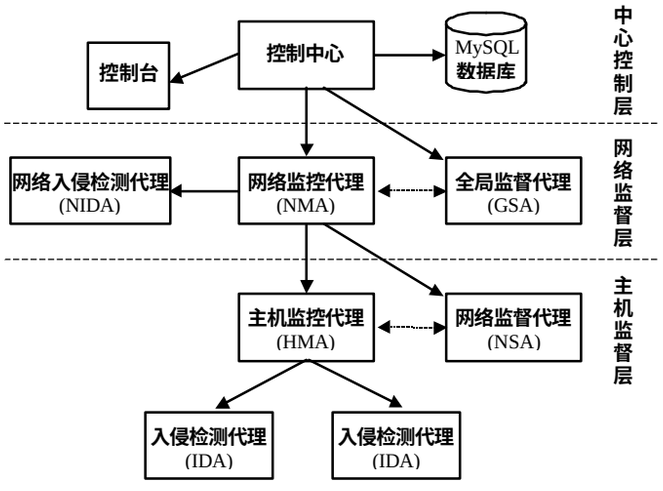

图 1 原型系统总体功能结构

原型系统总体功能结构如图1所示，整个系统为三层结构。其中控制中心对整个入侵检测系统的运行进行控制和管理，而控制中心的重要组成部分是数据处理模块；数据处理模块负责处理入侵警报、网络入侵信息、主机运行时数据和网络数据、对于主机运行时数据(主要指系统调用序列)、数据处理模块调用序列模式分析，从中挖掘规则。在挖掘过程中，使用遗传算法对挖掘数据进行优化，采用特征向量集代替特征属性变量集，利用遗传算法选择特征子集，从而高效地利用特征属性[2]。

2 遗传算法在入侵检测中的应用

2.1 遗传算法在基于模型入侵检测中的应用

应用遗传算法时，用一个个体代表一个可能的行为模型，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。适应度一般用预先定义范围的浮点数来表示，代表最优秀到最差的个体性能。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，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。

研究表明，个体的适应值可以取决于有多少攻击被正确检测和正常使用连接被误判为攻击，将适应度函数设计为[3]

$$F(x_i)=\frac{a}{A}-\frac{b}{B}$$

式中　x_i为某个个体；a为正确检测到的攻击数目；A为总的攻击数目；b为被误判为攻击的连接数；B为总的正常连接数。该适应度函数得出的适应度值在闭区间[−1,1]中，其中−1是最差的可能值，1是理想值。由此可见，a/A是检出率，b/B是误报率，高检出率低误报率使适应度函数值高，低检出率高误报率使适应度函数值低。

遗传算法生成的模型建立在解决入侵检测问题数据分析的新方法基础上。在模型的决策树上，每个结点数据被设计成拥有一个随机系数，数据与系数相乘成为判断该项数据记录是否代表攻击的确定性权重。这里的系数基于ERC(Ephemeral Random Constants)，是特定于数学建模的遗传算法生成的随机数，其微小变化也会导致进化变异产生。确定度为

$$C_i(x)=\sum_{j=1}^{n}R_{i,j}x_j$$

式中　C_i记录x是否被模型i识别为攻击的确定性大小；$R_{i,j}$为对应于结点特征x_j的随机系数；n为结点总数。取定一个随意的阈值，凡超过该阈值的确定度值所对应的记录则代表攻击。

2.2　遗传算法在一个新的入侵检测系统中的应用

在新系统的设计中，利用数据挖掘技术从系统日志、系统调用序列、网络流等大量数据中提取与安全相关的系统特征属性，为了高效地利用特征属性，采用特征向量集代替特征属性变量集，设计中采用遗传算法选择其特征子集，以降低入侵检测系统的负荷。

进行数据挖掘时，所选用的安全审计数据须具备以下特点：

1) 相对于正常的用户和系统行为，攻击事件的发生概率很小；

2) 在正常情况下所选用的安全审计数据非常稳定；

3) 攻击事件的发生会使安全审计数据的某些特征变量明显偏离正常值[4]。

特权程序一般都具有最高权限，因此特权程序一直是攻击者的主要目标。通过研究发现，对特权程序，系统调用序列较好地满足了数据挖掘对安全审计数据提出的要求，是理想的挖掘数据源。国外有关研究机构还提供了大量的有关系统调用序列的数据供IDS的研究者下载使用，基本上满足了完备性的要求。

系统调用序列检测的工作主要流程如下：

1) 准备训练数据集，该数据集中数据记录具有广泛的代表性，即具有较高的支持度；所有数据已经被准确标识为正常或异常，采用有关系统调用序列的数据作为分类器的训练数据集；

2) 用RIPPER算法分析训练数据集，提取特征属性，生成规则，

3) 基于所生成的规则，用滑动窗口法分析待检测系统调用序列。

为进一步提高IDS的性能，减少IDS组件对被保护系统的负荷，所设计的新入侵检测系统采用特征向量集代替特征属性变量集(短序列集)，在数据挖掘时产生了更简单、准确的入侵判别规则集。在此基础上进一步研究用特征向量子集代替特征向量集，采用遗传算法优化特征向量子集的选择过程，使IDS的性能得到进一步的提升。

在系统中，选取特征子集是考虑特征子集的选取是在入侵检测中提高机器学习算法性能的可行办法。特征子集的选取能提高学习算法的准确度，减少计算量，同时可以减少测试数据量，降低分类过程中的消耗等。进行特征子集选取，最重要的目标是提高入侵检测的准确率，减少分类运算等过程中的数据量。

在系统调用序列数据的挖掘过程中使用特征向量法，用特征向量的一位标识一个短序列，用挖掘算法能从特征向量集中找出检测入侵的规则来。由于短序列的数量较大，导致特征向量位数过大，特征向量集

也相应过大。为了更高效可行地使用数据挖掘算法，采用遗传算法对特征向量集进行优化，寻找特征子集，利于后续的数据挖掘。

在使用遗传算法的过程中，用特征向量的位数决定其个体的大小，随机构造50个二进制位串的个体，其中0、1代表该位置的短序列是否入选特征子集，如图2所示。在此基础上，进行遗传得到最优个体，该最优个体必然是0、1交替的位串，将其所有1所在位置进行分析，可以得到1所在位置代表的短序列集，即为寻找的特征子集。后续挖掘算法根据该特征子集中的短序列，对训练数据进行分类等挖掘工作。

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	…
0	0	1	0	1	0	1	0	1	0	1	0	1	0	0	1	…

图2 特征向量

采用标准交叉算子和变异算子，交叉概率取0.6，变异概率取0.001。遗传过程中，个体的选择比较复杂。因为这里是针对入侵检测进行的优化，所以在选择个体时，是将该个体代表的入选子集的短序列应用到数据分类算法(RIPPER)，该算法训练数据并应用规则得到测试数据，根据检测的性能来确定上述要选择的个体的适应度值。根据个体的适应度值就可以对其进行选择，继续遗传优化工作。

研究表明，个体的适应值可以取决于有多少攻击被正确检测和正常使用连接被误判为攻击，同时考虑个体中置1位的数目，本系统设计的适应度函数为

$$F(X_i)=\frac{[(a/A)-(b/B)]}{d\,m}$$

式中 X_i为某个个体，m为X_i中1的个数；d为m对于该适应度函数的相关系数，即高检出率低误报率使适应度函数值高，低检出率高误报率使适应度函数值低。个体中置1的位数越少，适应度值越大，这是出于寻找最小特征子集的考虑，其影响的强弱由相关系数d去控制。

2.3 算法基本步骤

本系统采用的遗传算法的基本步骤如下：

1) 设定进化代数$g=0$，生成包含n个个体的初始化群体$P(g)$；

2) 在该群体中对每个个体估值，计算各自适应度$f(x)$；

3) 根据个体适应度$f(x)$，从$P(g)$中选择两个个体作为父代(适应度值越大，选中的机会越大)，根据交叉概率，让选出的两个个体进行交叉产生新的后代(如果交叉概率为0，即不进行交叉，则后代就是父代的完全复制)，再根据变异概率，新生后代在各自基因座产生变异；重复上述步骤，产生新个体，将最后生成的个体形成新的群体$P(g+1)$；

4) 将新产生的群体$P(g+1)$作为后续进化操作所需的群体，令进化代数$g=g+1$；

5) 若终止条件满足，则算法结束，返回在当前群体中最好的个体，即最优解；

6) 若终止条件不满足，则跳至步骤2) 继续该遗传算法。

参 考 文 献

[1] 何险峰. 基于数据挖掘技术和智能体技术的入侵检测系统:[硕士学位论文][D]. 成都: 电子科技大学, 2003

[2] 黄 羽. 基于智能体技术的入侵检测系统及相关技术研究:[硕士学位论文][D]. 成都: 电子科技大学, 2003

[3] Jihoon Y, Vasant H. Feature subset selection using a genetic algorithm[C]. IEEE Intelligent System Special Issue: Feature Transformation Subset Selection, 1998. 44-49

[4] Salvatore J, Wenke L, Philip K C, *et al.* Data mining-based intrusion detectors: an Overview of the columbia IDS project[J]. SIGMOD Record, 2001, 30(4): 5-14

编 辑 徐培红