DOI:10.16208/j.issn1000-7024.2004.05.016

第25卷 第5期 计算机工程与设计 2004年5月
Vol.25 No.5 COMPUTER ENGINEERING AND DESIGN May. 2004

文章编号:1000-7024(2004)05-0708-05 中图分类号:TP393.4 文献标识码:A

基于数据挖掘的网络异常行为检测技术设计与实现

齐建东[1], 陶 兰[2], 孙总参[1]

(1. 中国农业大学 信息与电气工程学院, 北京 100083; 2. 深圳大学 信息工程学院, 广东 深圳, 518060)

摘 要: 既有的基于数据挖掘技术的入侵检测将研究重点放在误用检测上。提出了基于数据挖掘技术的网络异常检测方案, 并详细分析了核心模块的实现。首先使用静态关联规则挖掘算法和领域层面挖掘算法刻画系统的网络正常活动简档, 然后通过动态关联规则挖掘算法和领域层面挖掘算法输出表征对系统攻击行为的可疑规则集, 这些规则集结合从特征选择模块中提取网络行为特征作为分类器的输入, 以进一步降低误报率。在由DARAP1998入侵检测评估数据集上的实验证明了该方法的有效性。最后, 对数据挖掘技术在入侵检测领域中的既有研究工作做了总结。

关键词: 入侵检测; 异常检测; 数据挖掘; 审计记录;

Design and implementation of network anomaly behavior detection techniques based on data mining

QI Jian-dong[1], TAO Lan[2], SUN Zong-can[1]

(1. Information and Electric Engineering College, China Agricultural University, Beijing 100083, China; 2. Information and Engineering College, Shenzhen University, Shenzhen 518060, China)

Abstract: Data mining-based intrusion detection proposed before mainly focused on misuse detection. A data mining-based network anomaly detection technique were proposed. The implementation of kernel module was discussed in detail. Firstly, a combination of static mining algorithm on association rules and a domain level mining algorithm were used to profile the normal activity model of network. Secondly, a combination of a dynamic mining algorithm for association rules and the domain level algorithm, whose output consists of rules that characterize attacks to the system. These rules, along with a set of features extracted by a features selection module were used as the training set for a classifier for the purpose of lowering the false positive rate further. Experiment results on the DARAP 1998 intrusion detection evaluation dataset verified the effectiveness of this method. Finally, the work result about data mining-technique applied to intrusion detection system was summarized.

Key words: intrusion detection; anomaly detection; data mining; audit record

1 引 言

异常检测(anomaly detection)是当前入侵检测研究领域的热点, 普遍采用的方法是以大量的审计数据为背景来刻画系统或用户的正常使用模式, 建立正常活动模型, 然后通过检查当前活动和正常模型之间的偏离度, 来确认入侵行为。异常检测的优点很明显, 就是能够发现未知的攻击类型, 但较之误用检测(misuse detection), 误报率也很高。

数据挖掘技术已经在入侵检测中得到了应用, 哥伦比亚大学的Stolfo和Lee在这方面做了许多的开创性工作[1], 但他们的工作在检测策略上属误用检测范畴。本文提出一种异常检测方法, 具有如下两个特点: ①使用在线增量式挖掘: 系统并不是对TCP连接进行批量处理, 而是采用时间滑动窗口技术来查找该窗口内的可疑规则, 窗口值的大小根据攻击行为出现的频率高低来设定, 这样可以保证系统的实时性; ②执行异常检测, 首先使用静态挖掘算法和领域知识挖掘算法来描述网络正常活动简档, 之后在线挖掘过程中忽略那些与正常活动简档相符的规则集, 来寻找正常活动简档中未出现过的, 即未期望的规则, 并且通过分类器来进一步降低误报率。

论文内容结构如下: 第2部分描述了系统的组成及

收稿日期: 2003-03-18。

作者简介: 齐建东(1976-), 男, 内蒙古赤峰人, 博士研究生, 研究方向为入侵检测、数据挖掘; 陶兰(1956-), 女, 教授, 博导, 研究方向为计算机网络人工智能; 孙总参(1979-), 男, 硕士研究生, 研究方向为网络入侵检测、人工智能。

相关模块的划分，第3部核心模块的实现及相关技术分析，第4部分给出了在DARPA1998入侵检测评估数据集上的实验结果，第5部分总结了既有的数据挖掘技术在入侵检测研究领域应用的相关工作，最后做了简要的总结。

2 系统描述

我们使用结合分类算法的在线关联规则挖掘来确认可疑活动。关联规则[2]挖掘的目标是从数据库表中得出属性(features或attributes)之间的关联关系，形式如下：$X \rightarrow Y_{[c,s]}$，这里$X \cap Y=\Phi$，s是$X \cup Y$的支持度(表中同时包含X和Y的记录所占的百分数)，c是该规则的置信度，定义为$s_{x \cup y}/s_x$。关联规则挖掘算法中最重要也是最困难的部分是确定高于某一预定义阈值的频繁项集（一旦频繁项集确定了，便很容易根据置信度阈值得出形如$X \rightarrow Y_{[c,s]}$的关联规则)。在本文的技术实现中，旨在发现频繁项集(itemsets)，而不是它们之间的规则(rules)。贯穿全文我们将交叉使用这两个术语。

技术方案可描述如下：首先创建一个绝对不包含攻击数据的具有高支持度(大于预定义阈值)的关联规则数据库，该数据库中规则即为网络行为模式的正常活动简档。数据库中的项集按星期中的每一天(day/week)和每一天中的不同时间段(time/day)进行组织，这样可以进一步优化各规则与不同时期工作负载之间的专属程度。然后使用增量式、在线挖掘算法来检测当前具有较高支持度的规则，具体实现是在预定义大小为δ的时间窗口中使用动态挖掘算法，得到当前时间窗口中具有较高支持度的规则，并与数据库中已经存在的规则按星期中的每一天和每一天中的不同时间段进行比较。若当前规则已经存在数据库中，则忽略它，不占用存储资源来跟踪其支持度；反之，对该规则的支持度进行跟踪计数。当该规则的支持度超过某一阈值时，判定其为可疑规则，即对应的连接记录为可疑活动。对产生地可疑规则集实施两种方法：一是下钻到审计记录中找到产生该规则的原始数据；二是借助于一组向量参数来表征该规则集，并且把该向量作为决策树的输入。决策树进一步把可疑活动分类为正常事件、已知攻击种类或未知。

图1、图2描述了系统的总体结构，由两个阶段组成。阶段1为训练阶段，如图1所示。训练数据是以正常网络连接记录为背景，同时混有攻击类型数据，并对其加以标记。首先将不包含攻击数据的训练数据(纯净数据)，输入到一个模块中，该模块包含了用以挖掘关联规则的静态挖掘算法以及领域层面挖掘算法(后面将给出技术实现)，其输出就是网络行为的正常活动简档，即描述了没有攻击情况下的网络正常活动。然后将正常活动建档以及全部训练数据输入到一个模块中，该模块使用了动态挖掘算法和领域层面挖掘，其输出由表征攻击系统的规则组成，即可疑规则。最后将可疑规则连同从特征选择模块中提取的一组特征作为分类器(决策树)的训练数据。在使用系统检测入侵之前，训练阶段只实施一次。

阶段2为检测阶段，如图2所示。这里动态挖掘算法用来产生当前网络活动中可疑的规则集，连同由特征选择模块提取出的一组特征一起输入到在阶段1已经训练好的决策树中，该决策树把可疑规则分类为正常事件、攻击事件(同时标记其种类)或未知事件。当分类器把连接记录分为正常事件时，就把它们从警报中过滤掉，避免将其传给系统安全人员。对未知事件而言，其本质属性不能由分类器来精确指出。虽然未知类型事件不一定是未知类型的攻击，但在我们的实现中，视其为未知类型的攻击并在警报集合中包含进它们，传给系统安全人员以供进一步分析。本文下面部分，将进一步解释上述概念并给出系统的技术实现。

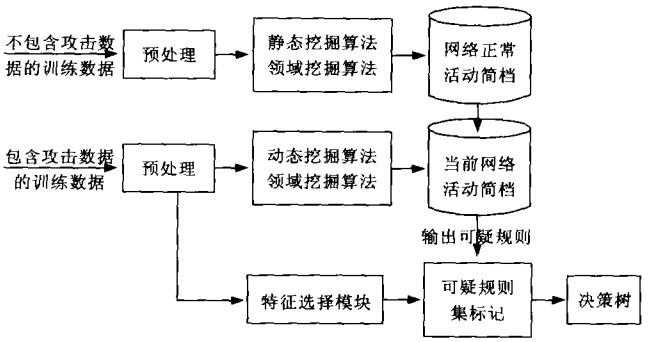

图1 训练阶段

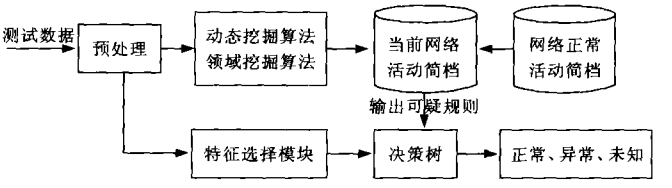

图2 检测阶段

3 技术实现

3.1 数据预处理

首先对TCP/IP连接记录进行预处理，预处理结果以如下机制保存在数据表中：R(T_s, Src.IP, Src.Port, Dst.IP, Dst.Port, FLAG)。其中T_s表示连接开始时间，Src.IP和Src.Port分别表示源IP地址和源端口号，Dst.IP和Dst.Port分别表示目的IP地址和目的端口号，FLAG用来描述TCP连接的状态，可以为open、reset、half open、finish等。关系R包含了关联规则挖掘地内容，这里讨论的关联规则挖掘比对一般的购物篮案例[2]的关联规则挖掘具有更强的约束性，例如规则左端不会出现两个不同的Src.IP值(不存在两个源端建立的连接)，但即便有诸多的约束条件，潜在的规则集也是巨大的，因为连接记录可能来自大量不同的源IP地址和源端口号。通过对攻击数据的分析，我们

选择了感兴趣的规则集模式，忽略了只包含源端IP地址而不包含目的端IP地址的规则集，因为它们缺乏对描述一个TCP/IP连接而言有价值的信息。另外，源端口号在TCP/IP连接建立时通常是随机选择的，并不体现该连接所对应的特定服务类型等有意义的信息，但这并不代表一个攻击行为总是符合该常规模式。我们在感兴趣的项目集中包含进了源端口号，目的就是要捕获所有可能的异常规则集，实验结果验证了这一假设。

3.2 静态关联规则挖掘

正常活动简档的建立是在训练阶段中完成的，对不包含任何攻击行为的训练数据进行离线处理，结果项目集中包含了刻画网络正常行为模式的活动简档。

在具体的算法实现中，将各感兴趣的项目集使用哈希表来存储，把各连接记录映射到哈希表i中，例如对存储Src.IP和Dst.IP的哈希表而言，$\Pi_{Src.IP,\ Dst.IP}(c)$代表了连接记录c到该哈希表的映射。然后通过查找和插入算法来完成全部连接记录到哈希表的映射，并过滤掉那些支持度低于预定义支持度的项目集。最终得到的结果为高于预定义支持度阈值的项目集，即网络正常活动简档。

3.3 动态关联规则挖掘

为发现可疑规则，我们使用滑动窗口方法来实现增量式在线关联规则挖掘，该过程称之为动态挖掘。这里只需对每个连接记录挖掘一次，这个特点使得该方法十分适合在线实时检测。滑动窗口的大小用Δw来表示，这个窗口大小既可表示Δw个连接，也可以作为是在Δw时间的连接记录。在具体的算法实现中，p_b表示指向滑动窗口中第一条连接记录的指针，$C[p_b]$代表p_b所指的连接记录，C_e是滑动窗口中最后一条连接记录。首先，检查当前连接记录C_e的映射$\Pi_i(C_e)$是否在对应的哈希表H_i中，若该映射包含在H_i中（即之前已经由动态算法找到），则增加其计数；否则在哈希表h_i中搜寻；若h_i中有，则忽略该映射，因为它属于正常事件；否则，将其存储在哈希表H_i中并将其计数（用$H_i.\Pi_i(C_e).Count$来表示）初始化为1，同时把该映射标记为INTERESTING。最后，计算标记有INTERESTING的规则的支持度$H_i.\Pi_i(C_e).Support$，当支持度超过阈值slevel时，该映射为可疑并输出。在遍历了C_e的每一个映射后，算法下滚窗口，减少每一个$C[p_b]$的映射支持度，把降为0的映射从相应的哈希表中删除，然后移动指针，得到一个新的连接记录C_e，开始新的挖掘过程。

3.4 领域层面挖掘

有时攻击行为是一些主机对另外一些主机发起的协同攻击，拒绝服务攻击就属于此类型。这时，Src.IP, Dst.IP形式的项目集可能会因为不具备足够的支持度而不会被系统标记为可疑规则。然而，如果我们把所有这种形式隶属于同一子网的项目集的支持度累计起来，其支持度大小将足够识别这类的可疑事件。

这种方法可以认为是对具有相同特征（隶属于一个子网）项目集的聚类，是通过累积每个单独的项目支持度来完成的。J. Han提出了多层面关联规则挖掘[3]，给出了使用自上而下的渐进挖掘方法来挖掘关联规则，尽管这里提出的概念也是多层关联规则挖掘，但是使用自下而上的方法来产生领域层面的规则。

对R(T_s, Src.IP, Src.Port, Dst.IP, Dst.Port, FLAG)机制而言，可以产生与IP相关属性的更高层的抽象。通常情况下，IP地址主机域的第一字节确认了该主机所属的子网，IP地址的前两个字节确认了该主机所在的网络号。例如，在领域层面挖掘中，从低到高定义了4层子网：Sub_1由IP地址的前三个字节确认；Sub_2由IP地址的前两个字节确认；Sub_3由IP地址的第一个字节确认；Sub_4是包含所有可能IP的子网的最高抽象层次。很明显，Sub_1是IP地址的第一层抽象，Sub_2是在Sub_1上的第一层抽象，依此类推。这样就对原感兴趣的项目集做了进一步的抽象。具体实现如下：设P(a)用来发现属性集a的下一层抽象。对一个项目集r=(a,b)而言，P(r)为在r上的项目集下一层抽象的父集和。于是P(r)由(P(a), b), (a, P(b)), (P(a), P(b))组成，这里P(a), P(b)分别是属性a,b的下一层抽象。例如对给定的(Src.IP, Dst.IP)形式的项目集而言，P(Src.IP, Dst.IP) = {Sub, Dst.IP; Src.IP, Sub; Sub, Sub}。我们开发了一个自顶向上的算法来挖掘领域层面关联规则，该算法很容易集成到动态挖掘算法中，因为它也使用一个哈希表来存储可疑项目集的支持度。

3.5 特征选择模块

特征选择本质上是个多窗口挖掘过程，动态挖掘采用的是单窗口大小，因为不同种类的攻击行为出现的频率不同，很难选择一个窗口大小的最优值来得到各种可疑规则。如果时间窗口值过大，算法会错失一些仅在短时间内发生的攻击类型；反之，将错失方式缓慢长时间进行的攻击行为。特征选择的目的就是克服单窗口大小的局限性。这里使用两个时间窗口值：一个是3秒窗口，用来捕捉那些在短时间内出现频率较高的规则集；另一个使用24小时窗口，用来捕捉那些持续时间长出现频率较低的规则集。在动态挖掘算法和领域挖掘算法上实施两个不同大小的时间窗口，从挖掘结果中提取出每秒平均连接率、从单一源IP地址对一组目的IP地址访问的毗邻指数等，这些特征用于后续的进一步分析。

3.6 警报分类模块

由关联规则挖掘算法得到的异常规则旨在用来指导检测工作。为尽可能地过滤掉误报，系统引入分类模块。使用一组属性来描述训练数据的规则（项目集），注意这里是对从训练数据中挖掘出来的项目集而不是针对原始训练数据。使用分类算法从训练数据中产生一个树结构形的分类器：其中叶子结点代表某一类属；树中的其它结点用来对实例的某个属性的测试，并且该结点的每一个后继分支对应于该属性的一个可能值。决策树用来对实例

图 3 DARPA1998 试验数据结果

星期		第 1 周					第 2 周					total
工作日		Mon	Tue	Wed	Thu	Fri	Mon	Tue	Wed	Thu	Fri	
#rules by MN	sp=0.01	1047	1440	1354	960	1063	1049	1355	1096	1228	993	11585
	sp=0.05	580	667	715	715	515	600	680	624	624	552	6084
	sp=0.1	476	553	555	457	447	538	525	549	538	489	5127
	sp=0.2	425	521	485	442	441	521	484	528	499	460	4806
#rules by DT	sp=0.01	6	10	4	57	16	79	15	6	17	16	226
	sp=0.05	6	10	4	41	16	79	12	6	16	15	202
	sp=0.1	6	10	4	41	13	78	9	6	17	12	195
	sp=0.2	6	9	4	41	9	78	8	5	13	11	184
#fp	sp=0.01	0	0	0	4	0	0	0	0	0	0	4
	sp=0.05	0	0	0	4	0	0	0	0	0	0	4
	sp=0.1	0	0	0	4	0	0	0	0	0	0	4
	sp=0.2	0	0	0	4	0	0	0	0	0	0	4
#det.	sp=0.01	3	4	3	6	6	10	7	4	9	10	62
	sp=0.05	3	4	3	6	6	10	6	4	9	10	61
	sp=0.1	3	4	3	6	6	10	5	4	9	10	61
	sp=0.2	3	3	3	6	5	10	2	4	8	9	56
#fn	sp=0.01	1	0	0	0	0	1	1	3	1	0	7
	sp=0.05	1	0	0	0	0	1	2	3	1	0	8
	sp=0.1	1	0	0	0	0	1	2	3	1	0	8
	sp=0.2	1	1	0	0	0	1	3	3	1	0	10

进行分类，从树的根结点开始，然后按给定实例的该属性值对应的树枝向下移动直至某一叶子结点。我们采用 C4.5[4]算法。由训练数据得到决策树后，使用它来对从在线挖掘算法中产生的可疑规则进行分类，即进一步将可疑活动分为正常事件、某类型攻击及未知。

通过对训练数据中的攻击进行研究，我们选取了一组属性，其中既有体现TCP/IP连接本质信息的属性，如服务类型、目的端口、连接持续时间等；也有在特征选择模块中提取出的属性，如每秒平均连接率、从单一源IP建立的一组目的IP的毗邻指数、从单一源IP连接的端口数目等。

4 试 验

我们使用的基于网络的安全审计数据来源于美国国防部高级计划研究署(DARPA)在1998年提供的用于入侵检测系统评估的数据[5]。这些数据包括了9个星期大约500万次会话，分为训练数据(training data)和测试数据(testing data)，其中训练数据包含了7个星期的正常数据和带标记的攻击数据，测试数据是由两个星期的正常数据和攻击数据组成。各星期中每一天的数据中都包含了TCPdump和BSM审计数据，我们的算法仅针对网络数据。DARAP 数据集目前是对入侵检测工具和方法评估最权威的数据集，有关该数据集更详细的信息见文献[5]。

图 3 给出了所检测到的攻击数目 (PROBE，DOS，PSSWD和DIC)、误报数目、漏报数目、当在动态挖掘算法中使用不同的决策支持度时作为决策树的输入的规则数目信息。标中的前两排给出了测试数据集的星期及星期中的每一天。total给出了每一行的总和；sp指的是动态挖掘算法中的支持度；#rules by MN指由挖掘算法得出的超出预定义支持度的并且不包含在正常活动简档的可疑关联规则数目；注意这里的可疑关联规则可能是正常活动也可能是入侵行为，为进一步加以确认，还要输入到决策树中进行分类；#rules by DT指决策树分类为异常的规则数目；#fp表示经决策树后处理后误报数目；#det 表示经由决策树判定的真正的攻击数目；#fn指经由决策树处理后的漏报数目。结果显示，当支持度下降时，输入到决策树的规则数目呈增加趋势；在动态挖掘算法中使用的支持度越低，就会检测到更多的攻击行为。图8还表明，决策树在降低由动态挖掘算法产生的误报上起了显著的作用。如果不采用分类引擎，由动态挖掘算法产生的所有异常规则都将被认为是攻击行为，通常每天超过400个，而仅有一小部分是和攻击相关的，所以误报率会相当高。此外，误报数目保持平稳并不受动态挖掘支持度的影响。所以，为检测到尽可能多的攻击，一个较好的办法是在动态挖掘算法中使用较低的支持度。

5 既有相关工作

5.1 关联规则

哥伦比亚大学的 Stolfo 和 Lee 在入侵检测研究中采用了数据挖掘技术，做了许多的开创性工作[1]。由于基本关联规则算法没有考虑任何领域问题，这样会产生大量的领域无关的规则。

Lee等人通过对基本关联规则挖掘算法进行扩展，来寻找系统调用以及用户活动在系统特征上的关联性。将从审计记录中挖掘的关联规则合并并且添加到最终规则集中来构建用户的正常行为模型 (当然训练数据必须绝对纯净，即不包含任何攻击数据)。规则的合并有两个条件：一是如果规则的左半部分LHS(规则的条件)和右半部分RHS(规则的结论)完全相同；二是支持度和置信度非常接近，即变化范围在一个非常小的值ε之间。具体算法设计参考文献[1]。

5.2 频繁情节规则

在入侵检测中考虑事件间的频繁模式是很重要的，因为有时入侵行为所产生的一系列事件中其单个事件是符合正常模式的，只有通过对这个系列事件进行完整分析才能断定是否有入侵行为发生。频繁情节(episodes)模式得到的是事件之间的频繁关系而不是事件内部属性之间的关系。Lee 和 Stolfo[1]扩展了基本频繁情节算法来计算频繁序列模式。共分两个阶段，首先使用轴心属性(axis

attributes，下文讨论)找到频繁关联规则，然后从这些关联规则中产生频繁序列模式。这里，时间窗口的选择至关重要，合适的窗口大小会挖掘出更充分的模式。

5.3 特征选择与精简

从审计数据中挖掘规则的一个问题就是某些属性(特征)对数据的描述而言是必须的，而另外一些则仅仅提供辅助信息。也就是说，在刻画数据的特征上，属性之间的"重要程度"是不同的。规则应该仅描述那些与"重要"属性相关的模式，这样的属性集称作 axis attributes(轴心属性或主因子属性)[1]。另外，还有一些属性，它们虽然和某特定攻击类型的特征(signature)有直接或间接关系，却可能和其它类型的攻击一点关系都没有。为降低计算代价以及实时响应入侵，必须确定能够充分描述数据的最小化属性集合。所以确定合适的属性集对IDS的性能而言是非常重要的。Lee 和 Stolfo[1]根据先验知识来选择轴心属性；Shi[6]使用遗传算法来自动选择适合特定攻击类型的属性集，试验表明这种方法比单纯依靠经验选择的方式在性能上有所提高。

5.4 模糊数据挖掘

审计记录中包含有量化特征，量化数据在挖掘过程中由支持度和置信度的阈值分隔在两个区间中。这种分区所带来的尖锐边界(sharp boundary)问题会对IDS的性能产生影响。为克服尖锐边界引发的问题，Bridges[6]使用数据挖掘算法(关联规则挖掘算法和频繁情节挖掘算法)与模糊逻辑相结合的方法开发IDS。具体地，Bridges把量化特征分为具有模糊隶属度值的几类。尽管试验结果令人满意，但模糊隶属度函数的参数是靠经验来选择的，这会引起错误报警。Shi[7]使用遗传算法来自动优化隶属度函数的参数，具体方法是首先定义有模糊函数参数序列组成的染色体，从一个随机的初始染色体种群开始，其中每个染色体都是一个可能的参数集合，然后使用适应度函数来为规则相似度设定优先级的高低。

5.5 多级近似 (level-wise) 挖掘

某些正常行为发生的频率较低，其支持度不足以高过预定的阈值，但如果在挖掘过程中采用很低的支持度，又会得到大量与频繁度高的服务类型相关的模式。为解决这一问题，Lee 和 Stolfo[1]提出了多级近似挖掘技术。其思想是首先找到那些与出现频率高的轴心属性值相关的模式，然后不断降低支持度阈值挖掘那些与出现频率低的轴心属性值相关的模式。在这个模式的挖掘过程中，限制那些"旧"的轴心属性的参与：候选项目集必须包括至少一个"新"(低频繁度)的轴心属性值。每次循环计算得到的模式或者是由所有"新"的轴心属性构成的或者是由"新"轴心属性与"旧"轴心属性一起构成。

5.6 对非标记数据的聚类挖掘

异常检测中正常活动简档的刻画，需要训练数据集十分完备并且数据绝对纯净，如果包含入侵行为的数据隐藏在训练数据中，算法将不会在将来的检测阶段将其识别出来。然而要获得带有标记或绝对纯净的数据是极其困难的，这不但是因为待处理的审计日志数据量非常巨大，而且对训练数据进行手工分类标记是十分繁琐且容易出错的工作。因此，Eskin 等[8]提出了一种非监督式异常检测算法，该算法使用一个简单的基于距离的测度将数据实例聚类成簇集，一旦数据归簇，就把一些小簇标记为异常实例。在 KDD CUP 99 数据集上的试验结果表明，与那些依赖于详细标记性数据的算法相比，尽管这种方法在误报率(False Positive)和漏报率(False Negative)上并不占优，但其优势也是明显的：不需要对数据进行预分类，不需要新的攻击类型的先验知识。

6 结 论

本文提出了网络行为的异常检测方法，给出了系统的总体框架，分析了各模块的功能，初步的试验结果验证了该方法的有效性，最后总结了数据挖掘技术在入侵检测领域中应用时所采用的技术。

参 考 文 献:

[1] Lee W, Stolfo S. Data mining approaches for intrusion detection[C]. San Antonio, TX: Proc. 7th USENIX Security Symposium(SECURITY'98). 1998.79-94.

[2] Agrawal R, Imielinski T, Swami A. Mining association rules between sets of items in large database[C]. Washington DC: Proceedings of the ACM SIGMOD Conference on Management of Data. 1993. 207-216.

[3] Han J, Fu Y. Discovery of multiple-level association rules from large databases[C]. Proceedings of the 21st Very Large Data Bases Conferences, Zurich, Switzerland, 1995.

[4] Ross Quinlan J. C4.5 programs for machine learning [Z]. Morgan Kaufmann, 1993.

[5] http://www.ll.mit.edu/IST/ideval/.

[6] Bridges S, Vaughn R. Fuzzy data mining and genetic algorithms applied to intrusion detection[C]. Proc. 23rd National Information Systems Security Conf. Baltimore, MA,2000.

[7] Shi F. Genetic algorithms for feature selection in an intrusion detection application, masters thesis [C]. Mississippi State University, Mississippi State, MS, 2000.

[8] Eleazar Eskin, Andrew Arnold, Michael Prerau,et al. A geometric framework for unsupervised anomaly detection: detecting intrusions in unlabled data[Z]. Data Mining for Securiyt Application(DMSA-2002), Kluwer, 2002.

更正：2004 年第 4 期第 564 页第一作者刘士清的单位更正为"中国科学院软件研究所"。