第 32 卷 第 19 期 计 算 机 工 程 2006 年 10 月
Vol.32 № 19 Computer Engineering October 2006
·博士论文· 文章编号：1000—3428(2006)19—0029—02 文献标识码：A 中图分类号：TP393

基于粗糙集理论的网络型入侵检测系统

张红梅[1,2]，王 勇[1,2]，王行愚[1]
(1. 华东理工大学信息学院，上海 200237；2. 桂林电子工业学院网络信息中心，桂林 541004)

摘 要：为解决目前大多数入侵检测产品或模型对未知攻击的检测都存在精度低或者虚警率高的问题，建立了一个基于网络的入侵检测实验平台，使用了多种新的攻击工具实施攻击；并在此基础上提取了网络连接的 29 项实时特征；应用粗糙集理论实现了一个网络连接的检测器。经实验表明，所选取的网络连接特征能较好地反映网络安全状况，粗糙集理论应用于多类分类问题和未知攻击的检测方面是有效的。
关键词：入侵检测系统；粗糙集；不可分辨关系；离散化；数据约简

Network-based Intrusion Detection System Using Rough Set

ZHANG Hongmei [1,2], WANG Yong [1,2], WANG Xingyu [1]
(1. School of Information Science and Engineering, East China University of Science & Technology, Shanghai 200237;
2. Network Information Center, Guilin University of Electronic Technology, Guilin 541004)

【Abstract】 Most of current products and models are poor at detecting novel attacks without an acceptable level of accuracy or false alarms. In order to figure out this problem, a network based intrusion detection system is established, and many up-to-date attack tools are used to attack the network. On the basis of the intrusion experiment, 29 variables are chosen as intrusion features to characterize the status of network connection. At the same time, the rough sets theory is exploited as a detector of network connection. The experimental results indicate that the features extracted from network connection are good indicators of the status of network and the rough sets theory is powerful in multi-class classification as well as effective in unknown attack detection.
【Key words】 Intrusion detection system (IDS); Rough sets; Indiscernibility; Discretization; Data reduction

由于网络攻击行为的不断加剧，建立有效的入侵检测系统来保护信息系统的安全变得越来越重要，也越来越具有挑战性。但目前大多数产品或模型对检测已知的攻击很有效，而对未知攻击的检测都存在精度低或者虚警率高的问题。

由于粗糙集理论在进行数据约简、数据特征提取时操作简便，并且在生成规则时不依赖任何数据之外的领域知识，生成的IF…THEN…规则易于理解等许多优点，使它比较适合于大规模的数据处理。已有一些研究工作将Rough Set理论应用到网络入侵检测系统中，但大多数使用的都是KDD Cup 99 IDS数据集，使用实测数据的研究比较少[1]。然而，随着计算机网络技术的发展，新的攻击手段可能使网络流量呈现全新的表征，原有数据集的局限性在所难免。因此，在新的网络环境下，使用各种不同的方法实现的入侵检测系统有重要的现实意义。

本文在入侵检测实验平台的基础上，使用了多种新攻击工具实施攻击，提取了网络连接的 29 项实时特征；并应用粗糙集理论实现了一个网络连接的检测器。

1 网络入侵实验介绍

本文中基于网络的入侵检测实验平台，与文献[1]中基于主机的入侵检测平台类似，但在攻击工具和特征采集方面有所区别。

1.1 系统的体系结构

入侵检测系统一般分为特征数据采集、特征预处理、入侵检测器、响应器等模块[1]。数据采集模块负责捕获网络信息，预处理完成数据规范化，本文采用了Rough sets 分类器作为入侵检测引擎，据预处理后的数据判断是否是入侵，处理结果送响应器。响应器根据具体情况采取相应的安全措施。

1.2 实验环境

整个实验网络由 1 台路由器、2 个集线器、3 台 Windows 主机、5 台 Linux 主机构成，网络拓扑如图 1 所示。

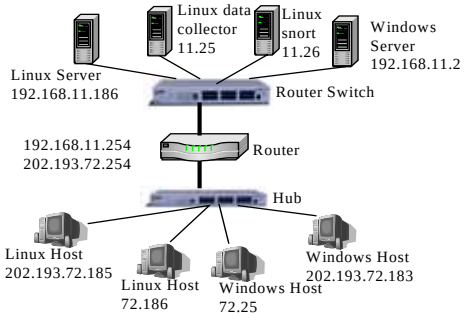

图 1 入侵检测实验环境

1.3 网络攻击

在实验中分别对 Linux 服务器和 Windows 服务器引入如下的攻击方式：综合扫描(probe)，拒绝服务(DoS)，远程攻击(r2l)，普通授权用户提升到特权用户的攻击(u2r)以及后门(backdoor)攻击，共 28 种攻击工具。

1.4 数据采集和预处理

通过进行了大量的网络攻击实验，实时地抓取数据包，

基金项目：国家自然科学基金资助项目(69974014)；教育部高校博士点基金资助项目(20040251010)
作者简介：张红梅(1970－)，女，副教授、博士生，主研方向：网络安全和智能信息处理；王 勇，副教授、博士；王行愚，教授、博士
收稿日期：2005-12-27 **E-mail：**hmzh2004@163.com

并根据这些数据包，为每个连接建立实时特征，从而根据这些特征判断连接是否正常。本文对每一连接，将从源至目的与从目的至源的包分开统计，从网络上收到第一个包起，至第 T ms，对该连接的包统计提取了 29 个特征项，其含义如表 1 所示，由于目的到源与源到目的类似，因此表 1 只列出源到目的 13 个特征项。

表 1 源到目的的入侵检测实时特征

Feature name	Discription	type
D_time	Sampling time/connection lasting time(μs)	
RF	Connection not finished：1；finished：0	0/1
Fn	Normal ended (to complete connection)	
	The features from source to destination	
Snum	Packets number sent within Tms	int
Sbyt	Bytes sent within Tms	int
Sfrag	Proportion of packets with “fragment” flag	real
Surg	Proportion of packets with “urgent” flag	real
Ssyn	Proportion of packets with “syn” flag	real
Sack	Proportion of packets with “ack” flag	real
Sfin	Proportion of packets with “fin” flag	real
Srst	Proportion of packets with “reset” flag	real
Spsh	Proportion of packets with “push” flag	real
Slen_mu	Mean value of packet length	real
Slen_sig	Standard deviation of packet length	real
Stime_mu	Mean value of packet network delay	real
Stime_sig	Standard deviation of packet network delay	real

2 基于粗糙集的检测器

在本文中，基于粗糙集的检测器功能主要是根据采集到的数据判断当前的网络连接是正常连接还是入侵。这个功能的实现包括 4 个步骤：(1) 数据预处理：包括删除重复记录，决策表补齐，数据离散化；(2) 求属性约简；(3) 根据值约简求出逻辑规则；(4) 使用规则进行分类。

2.1 基于粗糙集的检测器设计

在运用 Rough Set 理论从决策表中导出决策规则的过程中，样本数据的离散化和属性约简对结果的精度和效率影响较大，已经证明，最优离散化问题和最小属性约简计算问题是 NP 难问题。因此这两步的算法选择比较关键。

样本离散化方法一般包括：等距离划分，等频率划分，Naive Scaler，Semi Naive Scaler，布尔逻辑和基于属性重要性等算法。文献[2]使用Semi Naive Scaler算法，文献[3]使用等距离划分算法，而本文先用Nguyen 和 Skwron[4]提出的布尔逻辑算法，再用等频率划分算法。布尔逻辑算法的优点是在保持信息系统不可分辨关系不变的前提下，尽量以最少数目的断点集来把所有对象区分开。对于不需要保持不可分辨关系的属性，则留给等频率划分算法处理。

用于属性约简的算法有许多，如基于可分辨矩阵和逻辑运算的属性约简、基于特征选择的属性约简等算法。到目前为止，实际应用于大规模决策系统的最有效的算法还是Wroblewski提出的遗传算法[4]，该算法被广泛应用于各种粗糙集工具，如Rough Enough和Rosetta。文献[2,3]也采用了遗传算法进行属性约简，因此，本文也采用遗传算法来加快属性空间的搜索速度。

经过属性约简后，将多余的属性值删除即可完成值约简，然后从决策表中导出规则。若有多条规则的结论相互冲突，本文采用投票的方法解决。

2.2 样本数据离散化算法

由于输入数据表没有重复和缺失值，不需要做删除重复和补齐缺失。只需进行离散化。设S表示决策表，为了简单起见，假设所有条件属性A都是数值类型的，对于每一个α∈A，可以对它的属性值集合V_α排序，得到$V_a^1 < ... < V_a^i < V_a^{|V_a|}$，$C_\alpha$表示属性α自然产生的所有断点集合。

$X_a^i = \{x \in U \mid a(x) = V_a^i\}$；

$\Delta_a^i = \{v \in V_a \mid \exists x \in X_a^i \text{使} d(x) = v\}$；

$C_a = \{\frac{V_a^i + V_a^{i+1}}{2} \mid |\Delta_a^i| > 1 \text{ or } |\Delta_a^{i+1}| > 1 \text{ or } |\Delta_a^i| \neq |\Delta_a^{i+1}|\}$；

实际也就是，对于两个相邻的对象，在属性和决策值都不相同的情况下，选取两个属性值的平均值作为断点值，对于找不到断点集的属性，意味着该属性的不可分辨关系不需要保持，就暂时放着不处理，等本算法结束后用简单的等频率划分算法处理即可。

为了找到最小断点集，构造布尔函数 h。

$$h = \prod_{(x,y)} \sum_a \{\sum c^* \mid c \in C_a \; and \;\; a(x) \le (x) < c < a(y) \; and \;\; d(x) \neq d(y)$$

这里Σ表示布尔运算的累加，∏表示布尔运算的连乘，$d(x)$表示 x 对象在决策属性的取值。h 中的每个因子都是区分对象 x 和 y 的断点集的析取式，而这里面的每个析取式由每个属性α的断点析取式组成，只有那些能将 x 和 y 区分开的断点出现在析取式中。

2.3 入侵特征约简算法

本文采用遗传算法来计算最小入侵特征集，算法的主要运算过程如下：首先，随机产生一定数目的初始染色体，构成种群。其次，用适应度函数评价每个染色体的优劣。然后进行选择过程，选择优良的染色体，形成新的种群。对新种群进行交叉和变异操作，得到新种群。然后对新种群重复选择、交叉和变异操作，经过设定次数的迭代后，将最好的染色体作为解。适应度函数定义[4]如下。

$$f(B) = (1-\alpha) \times (|A| - |B|) / |B| + \alpha \times \min\{\varepsilon, [[S \; in \; S \mid S \cap B \neq \varnothing]] / |S|\}$$

其中 A 是条件属性的集合，B 是 A 的子集，它通过适应度函数驱动的进化搜索得到的最优属性约简集。

$S = \{M_D(i,j)_{n \times n} \mid M_D(i,j)_{n \times n} \neq \varnothing\}$；

其中ε是最小精度控制值；α是权重函数，适应度函数的第 1 部分，要求约简后的属性集合 B 的长度越短越好，第 2 部分要求 B 隶属于 S 的程度越大越好。

3 实验结果分析

实验分别对 Windows 和 Linux 主机实施攻击，采集了 3 010 组样本，其中 512 组是正常连接样本和 2 498 组攻击样本。将一半作为训练数据以得到检测模型，其中训练数据中包含 242 条正常样本，890 条扫描攻击样本；另一半作为测试数据测试得到的模型，其中 270 条正常样本，885 条扫描攻击样本。通过对 1 510 组训练样本进行训练，得到一个粗糙集分类器，其规则形式如：

Stime_mu([153648.00,*]) AND Dlen_sig ([757.19,1248.07])=> intrusion_id(0)

在“=>”左部的是条件属性，右部是分类属性，括号外的是属性名，里面是对应属性的取值。利用获取的规则，就可以对测试数据进行分类。当有多条规则相互冲突时，采用投票的方式决定最终的分类，同时还可得到分类的可信度。表 2 是将规则用于测试样本得到的混淆矩阵。

从表 2 可以看出，正常连接的检测精度为 94.07%，虚警率为 5.93%；扫描攻击的精度为 99.09%，而其余攻击类型的检测准确率较低。这主要是由于实验数据的分布不均匀引起的，因为从实验数据样本分布情况可以看出，正常连接和扫

（下转第 33 页）