维普资讯 http://www.cqvip.com

基于 Rough Set 理论的网络入侵检测系统研究*)

王旭仁[1,2] 许榕生[2] 张为群[3]

(首都师范大学信息工程学院 北京100037)[1] (中科院高能物理所计算中心 北京100039)[2]

(西南师范大学计算机科学系 重庆400715)[3]

摘 要 本文提出了一种基于 Rough set 理论(Rough Set Theory,RST)的网络入侵检测系统,用于监控网络的异常行为。该方法使用 Rough set 理论对网络连接数据提取检测规则模型。使用 Rough set 理论提取规则模型,能有效地处理数据挖掘方法中存在的不完整数据、数据的离散化等问题。实验表明,同其它方法相比,用 Rough set 理论建立的模型对 DoS 攻击的检测效果优于其它模型。

关键词 Rough set 理论,网络入侵检测系统,Semi Navie Scaler 算法,遗传算法

Network Intrusion Detection System Based on Rough Set Theory

WANG Xu-Ren[1,2] XU Rong-Sheng[2] ZHANG Wei-Qun[3]

(Information Engineering College of Capital Normal University,Beijing 100037)[1]

(Computing Center,Institute of High Energy Physics,CAS,Beijing 100039)[2]

(Computer Science Department of Southwest Normal University,Chongqing 400715)[3]

Abstract A network intrusion detection system based on Rough set theory has been developed,used in finding abnormal action in network. The approach mines detection rule models from network connection data. With rough set theory,several key problems can be solved,such as data completion,data discretization,and data reduction. Tests show that the network intrusion detection system based on Rough set theory has better performance in detecting DoS attacks than before.

Keywords Rough set,Network intrusion detection system,Semi navie scaler algorithm,Genetic algorithm

1 前言

网络安全入侵犯防范技术,例如用户认证(例如使用密码和生物测定学)、防火墙、信息保护(加密)已经被用于保护计算机系统的第一道防线。由于计算机网络越来越复杂,对系统信息的便利访问和控制的平衡也不可能让一个系统完全安全,因此入侵检测是保护计算机网络的另一种防护。

入侵检测(ID,Intrusion Detection)是"识别出那些未经授权而使用计算机系统以及那些具有合法访问权限,但是滥用(abuse or misuse)这种权限的人"[1]。进行入侵检测的软件与硬件的组合便是入侵检测系统(Intrusion Detection System,IDS)。入侵检测技术分为误用检测和异常检测两种,误用检测使用已知攻击的特征检测入侵,异常检测建立网络或者用户的正常使用模式,与此模式不一致的行为视为非正常行为。

自1980年 James P. Anderson 提出入侵检测的概念至今,经过20多年的发展,已涌现出上百种采用不同技术和方法,但大多数产品或模型存在缺乏精确性、缺乏检测新攻击能力、不便于分布式分析和协同工作、特征提取能力不足等缺陷。为了解决这些问题,把智能技术应用到入侵检测系统中,是近年来的研究热点。比较具有代表性的有:W. Lee 等人[2]在异常入侵检测系统中使用的关联分析、序列分析等技术建立用户正常使用网络的模型或规则;Forrest 等人根据人工免疫的原理,建立程序对系统的正常调用数据库("self"),作为检测异常调用的基础,以进程正常运行时产生的一定长度的系统调用短序列为模型来刻画进程正常运行状态[3];T. Lane[4]等建立了用户正常行为轮廓,应用机器学习中的实例学习(case study)方法和模式匹配检测攻击。但是,上述技术的特点是或者需要生成频繁项集,或者需要大数量的审计数据用于建立规则集,或者需要通过实验进行反复的特征选择,因此算法的代价比较大。而粗糙集理论可以在数据量的减少、数据特征提取上占有优势,几乎不需要人的干预,是一个代价较小的选择方案。一些已有的工作主要是利用 Rough set 理论对主机审计日志[5]和进程调用[6]进行分析。本文将讨论如何将 Rough set 理论应用到网络入侵检测系统中。

首先介绍 Rough set 的基本理论,然后介绍基于 Rough set 理论的网络入侵检测模型,最后是试验及讨论。

2 Rough set 基本理论

Rough set 理论假定知识是一种对对象进行分类的能力。而知识必须与具体或抽象世界的特定部分相关的各种分类模式联系在一起,这种特定部分称之为所讨论的全域或论域(universe)。

定义1 信息系统可用一个四元组来表示:$IS=\{U,A\cup$

*)基金项目:国家重点基础研究发展规划(973)项目(项目编号 G1999035806)。王旭仁 博士生,研究方向为网络安全和数据挖掘。许榕生 研究员,博导,研究方向为信息网络安全。张为群 教授,主要研究领域为人工智能,机器学习。

维普资讯 http://www.cqvip.com

$\{d\}, V, f\}, U \neq \varnothing$，表示对象(Object)的集合；$A$ 表示对象的全部属性(Attribute)；$V = \bigcup_{a \in A} V_a$ 是属性值组成的集合，V_a 表示属性 $a \in A$ 的值域；$f: U \times A \to V$ 是一个信息函数，指定 U 中每一个对象 x 的关于属性 a 的取值。

定义2 对于每个属性子集 $B \subseteq A$，定义二元关系 $IND(B)$：

$$IND(B) = \{(x,y) \mid (x,y) \in U^2, b \in B, s.t.\ b(x) \neq b(y)\}$$

称 $IND(B)$ 为不可辨识(不分明)关系。

显然 $IND(B)$ 是一个等价关系，且 $IND(B) = \bigcap_{Ib \in B} IND(\{b\})$；$\forall x \in U$，$x$ 基于不可辨识关系 $IND(B)$ 的等价类记作 $[x]_B$。

定义3 $T = \{U, A \cup \{d\}, V, f\}$ 是一个决策系统，其中 U，A，V，f 同定义1，$\{d\}$ 称之为决策属性，相应的 A 称之为条件属性。设 P 是 U^2 上的等价关系簇(集合)，若 $Q \subseteq P$ 有 $IND(Q) = IND(P)$，则称 Q 是 P 的一个约简(Reducts)，P 中所有必要的关系组成的集合称之为 P 的核(*Core*)，用 $CORE(P)$ 表示。

基于决策属性 $D = \{d\}$ 的可辨识矩阵定义为：

$$M_D = M_D(i,j)_{n \times n} = \begin{cases} \{c_k \mid c_k \in C \wedge c_k(x_i) \neq c_k(x_j)\}, & d(x_i) \neq d(x_j) \\ \phi, & d(x_i) = d(x_j) \end{cases}$$

已经证明最小约简是 NP-Hard 问题，但是可以通过智能技术得到最小约简。

3 基于 Rough set 理论的网络入侵检测系统

3.1 系统结构

从图1可以看出，基于 Rough set 的网络入侵检测系统包含两个阶段：模式生成和模式检测阶段。在模式生成过程中，网络数据收集模块用来收集网络连接数据，用作规则生成的原始数据；而数据选择模块是根据数据分析的任务，指从收集的数据中选择出目标数据，包括数据的维数、属性、数据的类型等。Rough set 理论应用在数据预处理模块和知识约简模块中。模式检测阶段使用生成的检测规则，对当前数据进行检测并对异常行为产生报警。

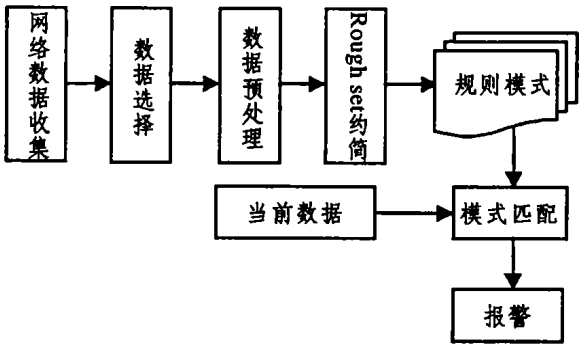

图1 基于 Rough set 的网络入侵检测系统

3.2 数据预处理

数据预处理包括不完整数据的补齐和数值属性的离散化。

在很多情况下，待处理的数据集并不是一个完备的数据表。数据补齐算法可供选择的很多：直接删除不完整的对象；根据统计方法对遗漏属性值进行估计补充，例如平均补齐算法、条件平均补齐算法、组合补齐算法、条件组合补齐算法等；根据 Rough 集理论中数据不可分辨关系来对不完备的数据进行补齐处理等。

Rough set 理论是基于集合论的基础上提出来的，所以对数据中连续属性要进行离散化处理，包括对连续数据的离散化和对离散数据的离散化：连续值(如浮点型数据表达)在处理前必须进行离散化处理；将离散数据进行合并(抽象)得到更高抽象层次的离散值。最优离散化问题(离散的切点数最少)已被证明是 NP-Hard 问题。在选择离散化算法时，要考虑离散化后的数据保持原有的不可分辨关系，主要有基于可辨识矩阵和逻辑运算的离散化方法、贝叶斯方法、信息熵方法、聚类法等。

3.3 知识约简

知识约简主要有两大分支，一是使用分类之间的可辨识性来进行属性选择，在属性约简的过程中没有回馈信息，例如借助可辨识矩阵与数学逻辑运算求最佳属性的方法[7]；把属性的频率函数作为属性重要性的度量，提出了基于频率的属性约简算法[8]；动态约简的方法[9]。二是使分类回馈进行属性选择，如基于互信息的知识约简的算法 MIBARK[10]；基于 Rough set 理论对数据进行约简后，再使用神经网络对约简的结果进行过滤、去除噪音，得到更加精简的约简，加强了规则的鲁棒性，从而避免了直接使用神经网络提取规则的困难[11]；S. Vinterbo 和 A. Øhrn[12]提出了使用遗传算法搜索属性空间，加快搜索速度。

4 试验和讨论

4.1 数据收集

在网络安全领域建立入侵检测系统，需要大量的数据来生成模型。获取数据有两种途径：一是自己建立模拟网络，收集相关数据，在收集数据时会受到数据隐私、安全等问题的困扰；二是使用以前的数据集，优点在于便于和以前的工作进行比较，并且避免了上述提到的问题。

选择 KDD CUP 99 Data Sets 作为数据源，每个网络连接记录有41个属性值，相当于信息系统中的条件属性。条件属性有四种类型：基本连接属性、连接内容属性、基于连接的时间和基于主机的属性。基本属性有每个网络连接的基本特征，包括：协议类型、连接持续时间、连接的客户端和服务器各自发出的数据长度等；内容属性，是根据 U2R 和 R2L 攻击的领域知识进行搜集的，包括：登录失败的次数、使用 root 命令的次数、访问根的次数、创建文件的次数等；基于时间的属性以“2秒”作为时间窗口，观察时间窗口内各种连接和当前连接的关系，包括：2秒内和当前连接具有同样目的 IP 的连接数，2秒内到达同一目的主机的连接中 SYN 出错率、REJ 出错率、相同服务率、不同服务率，2秒内和当前连接具有相同服务的连接数等；基于主机的属性有：描述前100个连接和当前连接线的关系，属性类似于基于时间的连接。每一条网络连接记录都预先进行了标记(即分类)：正常连接、DoS(Denial of Service)攻击、R2L(Remote to Local)攻击、U2R(User to Root)攻击，为处理方便起见，分别用0、1、2、3、4表示。每种攻击又有不同的变种。

实验中的训练数据选取了41个属性中的36个，其中7个基本属性，11个连接的内容属性，18个连接的基于主机和时间的属性。没有使用的5个属性由于它们在实验数据中取值相同或者全都为0，因此忽略。训练数据中有13,107条连接记录，而测试数据中有26,214条连接记录，数据连接的分布如表1所示。且有8种攻击864个连接只在测试数据中出现，其中 probe 类型的连接有250个，DoS 的连接有160个，U2R 的连接有1个，R2L 的连接有453个，用这些测试数据中的未知攻击来检验

维普资讯 http://www.cqvip.com

模型的检测效果。

表1 实验数据分布

	训练数据	测试数据
类型	Percent(%)	Percent(%)
normal	59.952698	9.727627
probe	1.213092	1.224537
DoS	38.735027	87.308311
U2R	0.015259	0.003815
R2L	0.083925	1.735714

4.2 数据预处理

由于数据源中的数据没有不完全数据，因此没有进行数据补齐的工作。

在实验数据中有较多的数值属性(30个)，使用 Semi Navie Scaler 算法进行离散化。断点的选取完全由数据本身的信息决定，在离散化后保留了数据原有的不可分辨关系。具体的算法如下：

对于每一个属性 $a\in A$，断点集合 $C_a=\phi$，执行以下操作：

1)对值域 V_a 进行排序，使得相邻的对象(记录)x_i、x_{i+1}有：$v_a^i\leqslant v_a^{i+1}$，v_a^i，$v_a^{i+1}\in V_a$。

2)计算 x_i、x_{i+1}所属的等价类对应的决策中出现频率最多的决策值的集合 D_i，D_{i+1}：

$D_i=\{v\in V_d|v=\arg\max_{v'}|\{x\in[x_i]_a|d(x)=v'\}|\}$；

$D_{i+1}=\{v\in V_d|v=\arg\max_{v'}|\{x\in[x_j]_a|d(x)=v'\}|\}$；

3)if $((D_i\subseteq D_{i+1})$ or $(D_{i+1}\subseteq D_i)$

$C_a=C_a\cup\{(v_a^i+v_a^{i+1})/2\}$；

4.3 知识约简算法

在进行约简时，采用遗传算法(GA)对对象空间进行搜索，加快知识约简的速度。适应度函数定义如下：

$$f(B)=(1-\rho)\times\frac{|C|-|B|}{|C|}+\rho\times\min\{r,\frac{|[S\text{ in }S|S\cap B\neq\phi]|}{|S|}\}$$

其中 ρ 是权重函数，C 是条件属性集合，$S=\{M_D(i,j)_{n\times n}|M_D(i,j)_{n\times n}\neq\phi\}$，$r$ 是最小精度控制值，适应度函数的第一部分，要求约简后的属性集合 B 的长度越短越好，第二部分要求 B 隶属于 S 的程度越大越好。算法如下：

1)随机选取初始群体 P，循环作以下操作：

2)对 P 进行选择(selection)得到 P_1，P_2，P_3；

3)对 P_1进行交叉(crossover)得到 Q_1；

4)对 P_2进行变异(mutation)得到 Q_2；

5)对 P_3进行倒位(Inversion)得到 Q_3；

6)根据2)3)4)5)的结果计算出新的群体 P；

7)如果 P 中的基因的适应度不再增加或 P 达到指定代数，退出；

8)否则，转到步骤2)；

4.4 实验结果和分析

实验所用的分析工具是由挪威科技大学(Norwegian University of Science and Technology)[12]计算机和信息科学系的知识系统研究小组开发的软件包 ROSSETA——一个图形界面下的 Rough set 分析软件完成的。实验产生的约简规则左部条件属性最多为5个，规则形式如下所示：

protocol_type (tcp) AND flag (SF) AND dst_bytes ([*,3)) AND num_failed_logins([*,1)) AND dst_host_same_srv_rate([0.09,*)) ⇒ class(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，里面是对应的属性值，它表示当规则左部的情形出现时，得到的分类是正常连接。利用获取的规则，就可以对测试数据进行分类，所得到的混淆矩阵(confusion matrix)如表2所示。

表2 测试数据分类混淆矩阵

		预测分类					
	Type	0	1	2	3	4	%correct
实际分类	0	2113	0	437	0	0	0.83
	1	74	1	246	0	0	0.003
	2	24	11	22852	0	0	0.99
	3	1	0	0	0	0	0.0
	4	25	1	429	0	0	0.0

从表2中可以看出，正常连接和 DoS 攻击的连接检测准确率较高，DoS 攻击的连接检测准确率优于 KDD Cup99分析的最好结果97.26%。由于 DoS 攻击会对网络造成很大的威胁，因此试验的检测结果很有意义。而其他的攻击检测准确率很低。这可能由以下原因引起：1)试验数据分布不太合理：probing，U2R，R2L 等攻击类型的比例太少，例如在训练数据和测试数据中三种攻击所占比例都不到1.3%，U2R 连接只有2个，这对规则的有效提取和攻击检测可能产生很大影响。在全部 KDD CUP99的实验数据中都存在着数据分布不平衡的问题；2)在 KDD CUP99获胜者的分析中，对 U2R 的检测准确率最高不到14%，R2L 不到9%，这说明检测结果不理想还和 KDD CUP99收集的属性不足有关。在文[4～6]中通过对主机审计日志和进程调用等详细分析，能够获得对 U2R、R2L 类型攻击的较好的检测结果，在上述这些工作中，都收集了基于内容和主机的更多的属性，例如针对缓冲区溢出攻击所收集的记录长度、字符类别、间隔时间属性[5]，对正常系统调用的详细命令序列的分析来发现异常行为[4,6]等，这些都是 KDD CUP99数据集所不具备的属性。

参考文献

1 Mukherjee B, Heberlein T L, Levitt K N. Network intrusion detection. IEEE Network, 1994, 8(3): 26～41

2 Lee W, Stolfo S J, Mok K. Data mining in workflow environments: Experiences in intrusion detection. In: Proc. of the 1999 Conf. on Knowledge Discovery and Data Mining (KDD99), 1999

3 Forrest S, Perelson A S, Allen L. Self-nonself discrimination in a computer. In: Proc. of the 1994 IEEE Symposium on Research in Security an Privacy, 1994

4 Lane T, Brodley C E. Detecting the Abnoraml: Matchine Learning in Computer Security: [Technical Report ECE-97-1]. 1997

5 冯力，彭勤科，管晓宏. 基于粗糙集理论的安全日志分析模型. 计算机工程，2002，28(11)

6 蔡忠闽，管晓宏，邵萍. 基于粗糙集理论的入侵检测新方法. 计算机学报，2003，26(3)

7 常犁云，王国胤，等. 一种基于 Rough set 理论的属性约简及其规则提取方法. 计算机研究与发展，1999，10(11)：1206～1211

8 王珏，王任，苗夺谦，等. 基于 Rough Set 理论的数据浓缩. 计算机学报，1998，21 (5)：393～399

9 Bazan J G, Skowron A, Synak P. Dynamic reducts as a tool for extracting laws from decision tables. In: Proc. Intl. Symposium on Methodologies for Intelligent Systems, volume 869 of LectureNotes in Artificial Intelligence, Springer-Verlag, 1994. 346～355

10 苗夺谦，胡桂荣. 知识约简的一种启发式算法. 计算机研究与发展，1999，36(6)：681～684

11 Li Renpu, Wang Zheng-ou. Mining classification rules using rough sets and neural networks. European Journal of Operational Research, Sept. 2003

12 Vinterbo S, Øhrn A. Minimal approximate hitting sets and rule templates. Intl. Journal of Approximate Reasoning, 2000, 25(2): 123～143