ISSN 1000-0054 清华大学学报 (自然科学版) 2008年 第 48卷 第 7期 w26
CN 11-2223/N J Tsinghua Univ (Sci & Tech), 2008, Vol. 48, No. 7 http://qhxbw.chinajournal.net.cn

基于粗糙集的自适应入侵检测算法

赵曦滨[1,2], 井然哲[1,2], 顾 明[1]

(1. 清华大学 软件学院, 北京 100084; 2. 国家企业信息化应用支撑软件工程技术研究中心 苏州制造业信息化研究室, 苏州 215104)

摘 要: 为了提高入侵检测系统的检测率,降低错检率,在分析现有入侵检测方法基础上提出一种基于粗糙集的入侵检测算法,将粗糙集算法和入侵检测技术结合起来实现系统的安全检测。对收集到的入侵数据进行预处理、数据离散化,属性约简,并依据生成的检测规则来分析入侵数据。实验结果表明: 与基于 BP(back propagation)神经网络和支持向量机的入侵检测算法比较,该算法的检测率提高 10% 左右,能很好地为信息系统提供入侵检测服务。

关键词: 粗糙集; 计算机网络安全; 知识约简; 入侵检测; 检测规则

中图分类号: TP 393.08 文献标识码: A

文章编号: 1000-0054(2008)07-1165-04

Adaptive intrusion detection algorithm based on rough sets

ZHAO Xibin[1,2], JING Ranzhe[1,2], GU Ming[1]

(1. School of Software, Tsinghua University, Beijing 100084, China;
2. Suzhou Manufacturing Industry Information Research Center, National Engineering Research Center for Enterprise Information Software, Suzhou 215104, China)

Abstract Intrusion detection systems are automatic system which recognize intrusions to computers or computer network systems. Existing security detection systems have many problems such as wrong detection of intrusions, missed intrusions, poor real-time performance. An intrusion detection algorithm was developed by combining a rough set algorithm with intrusion detection technology for security detection. The algorithm includes data preconditioning, data discretization, attribute reduction, production of detection rules, and finally analysis of intrusion data with these rules. Test results show that the intrusion detection algorithm is more efficient than algorithms based on BP neural networks and vector machines; thereby, improving the detection ratio by about 10% and reducing the wrong detection ratio. The system provides detection service effective for information systems.

Key words rough sets; computer network security; attribute reduction; intrusion detection; detection rules

随着网络技术的发展和网络规模的扩大,网络安全问题逐渐显露,建立有效的入侵检测系统以保护信息系统的安全变得越来越重要。入侵检测系统(intrusion detection system, IDS)是对计算机或计算机网络系统中的攻击行为进行检测的自动系统[1-2]。入侵检测由传统电子数据处理、安全审计以及统计技术发展而来并在很多的安全系统中得到应用。

目前,入侵检测技术得到了很大的发展,但仍存在误报和漏检、缺少自我防御功能、时间性差和协调性低等问题[3]。很多学者提出了改进的方法: 文[4]把入侵检测看作是区分正常和非正常的过程,提出了基于免疫模型的入侵检测技术; 文[5]把支持向量机技术应用于入侵检测系统,该方法避免了基于传统机器学习的局限性,保证了较强的推广能力; 文[6]利用神经网络来提取特征和分类; 文[7]提出一种强化规则学习的入侵检测方法,将规则学习算法应用到入侵检测模型中,有效降低了误报率; 文[8]从数据挖掘技术角度探讨了入侵检测的实现问题。

但是,将数据挖掘应用于入侵检测必须依托于大量的数据,利用支持向量机方法训练时间较长,建立入侵检测模型较困难,而采用基于粗糙集的方法进行检测知识约简,能够发现潜在的、有效的检测规则,使系统的检测性能得到了进一步的提高。本文将粗糙集理论和入侵检测技术相结合,通过知识约简产生最小分类检测规则,判断数据和行为的正常和异常情况,从而有效提高检测率,降低错检率,以更好地为入侵检测系统服务。

收稿日期: 2007-04-24
基金项目: 国家自然科学基金资助项目 (90412007)
作者简介: 赵曦滨(1973-), 男(汉), 江苏, 讲师。
E-mail: zxb@tsinghua.edu.cn

1 粗糙集理论基础

粗糙集[9]是波兰华沙理工大学 Pawlak 教授于1982年提出的一种处理不完备信息的方法。它不需要任何先验信息,能够有效分析和处理不完备、不一致、不精确的数据。通过对大量数据进行分析,根据论域中的两个等价关系的依赖关系来剔除相容信息,并抽取潜在有价值的规则知识。该方法已经在知识获取、规则提取、机器学习、决策分析、模式识别、数据挖掘等领域获得了广泛的应用[10],非常适合安全规则的学习和发现。本文尝试采用粗糙集产生的规则来有效检测用户的数据和发出的行为。

定义 1 给定集合 U 和等价关系集合 R,在等价关系集合 R 下对数据集合 U 的划分,称为知识,记为 U/R。

定义 2 一个给定的知识库是一个关系系统 $K=(U,R)$,U 为论域,R 是 U 上等价关系的一个族集。

令 $X\subseteq U$,R 为 U 上的一个等价关系。当 X 能表达成某些 R 基本范畴的并时,称 X 是 R 可定义的,否则称 X 为 R 不可定义的。R 可定义集也称作 R 精确集,而 R 不可定义集也称为 R 非精确集或 R 粗糙集。对于粗糙集可以近似地定义,使用两个精确集,即粗糙集的上近似和下近似来描述。

定义 3 X 的 R 下近似:

$$\underline{R}X=\bigcup\{Y\in U/R \mid Y\subseteq X\}.$$

定义 4 X 的 R 上近似:

$$\overline{R}X=\bigcup\{Y\in U/R \mid Y\cap X\neq\Phi\}.$$

知识约简是粗糙集中的核心内容之一,所谓知识约简,就是在保持知识库分类能力不变的情况下,删除其中不相关或不重要的知识。

定义 5 设 $Q\subseteq P$,若 Q 是独立的,且 $\mathrm{ind}(Q)=\mathrm{ind}(P)$,则称 Q 为 P 的一个约简。P 中所有必要关系组成的集合称为 P 的核,记作 $\mathrm{core}(P)$。核与约简有如下关系:$\mathrm{core}(P)=\bigcap \mathrm{red}(P)$。

定义 6 一个知识表达系统是一个四元组 $S=(U,A,V,f)$,其中,U:对象的非空有限集合,称为论域;A:属性的非空有限集合;$V:\bigcup_{a\in A}V_a$,V_a 是属性 a 的值域;f:$U\times A\to V$ 是一个信息函数,它为每个对象的每个属性赋予一个信息值,即:

$$a\in A,\quad x\in U,\quad f(x,a)\in V_a.$$

决策表是一类特殊而重要的知识表达系统。多数决策问题都可以用决策表形式来表达,这一工具在决策应用中起着非常重要的作用。

定义 7 设 $S=(U,A,V,f)$ 为一知识表达系统,$A=C\cup D$,$C\cap D=\Phi$,C 称为条件属性集,D 称为决策属性集,具有条件属性和决策属性的知识表达系统称为决策表。

定义 8 (a,v),(a 表示属性值,v 表示属性的取值)是原子公式;原子公式也是公式。

定义 9 如果 A 和 B 是公式,那么 $\neg A$,$A\cup B$,$A\cap B$,(A),$A\to B$ 都是公式。

定义 10 只有按定义 4.8和定义 4.9组成的式子是公式。

定义 11 公式 $A\to B$ 的逻辑含义成为决策规则,A 成为规则前件,B 称为规则后件,它们表达一种因果关系。其中公式 A 中所包含的原子公式中只有决策表中的条件属性,B 中所包含的原子公式中只有决策表中的结果属性。

2 粗糙集入侵检测模型

本文提出的基于粗糙集的入侵检测模型,如图 1 所示,先对收集的数据进行数据预处理,选择训练样本,权值离散化,进行决策表属性的约简,产生约简输出规则,从而构造安全系统的规则库,产生入侵检测器,建立初始的检测模型,并在以后系统运行中逐步完善和改进模型,以达到最好的检测效果。

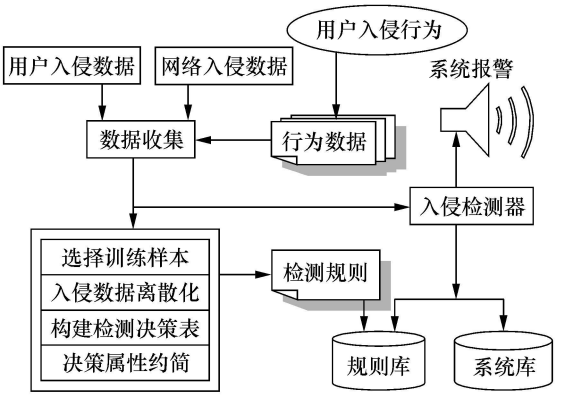

图 1 基于粗糙集安全入侵检测模型

由检测模型可以清楚地看到,入侵检测算法主要涉及以下几个基本问题。

1) 入侵数据离散化。入侵检测系统待分析的数据包括网络数据和主机数据,网络数据包的分析是当前入侵检测研究中的一个侧重点。与主机日志数据相比,网络数据更加复杂多样,而基于网络攻击的检测难度也大大增加。为了提高检测的效果,需要对采集到的大量数据进行离散化,离散化方法采用等频率划分的方法。

2) 入侵属性约简。对收集的数据集进行属性规整，对入侵检测属性进行约简(即在关系数据库中删除重复的行)，去除冗余入侵属性。

3) 入侵检测规则的产生。经过属性约简后，将多余的属性值删除即可完成值约简，构建决策表，然后从决策表中导出规则。对形成的规则进行检测和核实，放到规则库，安全检测器根据规则库里的规则进行数据和行为的入侵检测。

3　基于粗糙集入侵检测算法

入侵检测算法可分为 2个阶段进行，检测规则生成和检测模型的学习更新，基于粗糙集入侵检测算法思想描述如下：

算法　基于粗糙集的入侵检测算法

输入　系统采集到的用户使用数据和行为数据集 $\boldsymbol{S}=\{s_i \mid s_i \in \text{Userdata or Actdata}\}$，表示为粗糙集中的属性集 $\boldsymbol{A}$

输出　数据和行为的检测规则 $\boldsymbol{R}=\{r \mid W_i => C_i\}$，然后加入到安全规则库 $\boldsymbol{K}=\{k_i \mid k_i \in \text{IDR-Base}\}$。

阶段I　设置入侵检测规则

步骤 **1**　收集入侵数据集 $\boldsymbol{S}$，进行数据预处理，删去重复和多余的属性 a_0，$\boldsymbol{A}=\boldsymbol{A}-a_0$ 补齐决策表 ST。

步骤 **2**　For $i=1$ to n，do 对于任意属性 $a \in \boldsymbol{A}$，根据属性值 Sa，进行从小到大(Sort(w))，根据属性权值最大值和最小值，预先给定参数 k，得到断点集 Con(i)，$i=1$ to k。

步骤 **3**　利用权值离散化的属性构建决策表 ST，以经过约简的属性集合 $\boldsymbol{S}'$为条件属性，所有安全规则作为决策属性集。

步骤 **4**　进行属性约简，For $i=1$ to n，do 对于属性集合 $\boldsymbol{S}$中的各个属性 a_i 进行属性约简检测 Card(a_i)，set reduce-set=$\boldsymbol{S}$，if (Card($\boldsymbol{S}$)=Card($\boldsymbol{S}-a_i$)，then delete(a_i)，end do。

步骤 **5**　化简和分析决策表，剔除多余和不合理的规则 r_0，使得 $\boldsymbol{R}=\boldsymbol{R}-r_0$ 产生安全检测规则集合 $\boldsymbol{R}=\{r \mid W_i => C_i\}$。

阶段II　入侵检测分析

步骤 **6**　将化简得到的安全检测规则交给安全检测器进行数据测试和分析，合理则 $\boldsymbol{K}=\boldsymbol{K}+\boldsymbol{R}$

步骤 **7**　选取合适的训练样本进行多次训练，完善安全检测模型并更新检测器，直到满足一定的误用率和虚警率为止。

If(Check(r_i)==true or GetNewRule())

Then $\boldsymbol{K}=\boldsymbol{K}+r_i$;

Else if(Check(r_i)==false or DeleteRule())

Then $\boldsymbol{K}=\boldsymbol{K}-r_i$;

End if

步骤 **8**　算法结束。

4　算法的评价和实验分析

4.1　实验环境

为了证明本文算法的有效性，实验中采集了各种不同的检测数据，描述了每个网络连接的本质特征：持续时间，协议类型，服务类型，服务端和客户端发出的数据长度，连接状态标志等；连接的内容属性描述了每个网络连接的行为：登录失败的次数，使用 root 命令的次数，是否成功登录系统，访问存取控制的文件次数等。实验环境操作系统为 Windows NT 2000，Inter(R) Pentium(R) CPU 2.4 GHz，内存：1.0 GB。程序执行环境采用 Visual C++ 6.0和执行粗糙集算法的软件 ROSE2。

4.2　实验结果

将收集到的全部入侵数据进行实验，经过数据预处理和属性权值离散化，得到训练数据 4450条，测试数据 2045条，进行基于粗糙集的入侵检测算法处理，经过多次的测试，得到了如下的结果(见表 1)。

表 **1**　检测率和错检率测试数据

%

攻击方式	检测率	错检率
正常数据	95.22	4.78
Dos攻击	87.16	12.24
U2R攻击	79.53	12.13
R2L攻击	80.82	13.07
Probe攻击	83.19	9.69

从表格的数据统计可以看出，通过基于粗糙集的检测方法进行网络入侵检测，算法在检测率和错检率上得到了改进，很好地满足了系统的安全检测需求。但由于知识约简的原因，可能将一些属性删除，使得数据的完整性受到影响，有信息丢失的情况，并且在表格中体现出某个类别的检测率偏低而错检率偏高。但是从其他攻击方式的检测率和错检率来看，算法的效率和有效性还是得到了很大的改进。为了进一步检验算法的有效性，同其他入侵检测方法进行了对比实验，比较的方法有基于数据挖掘

方法(DM)、支持向量机方法(SVM)和 BP神经网络方法(BP)等. 得到如下实验结果(如图 2所示).

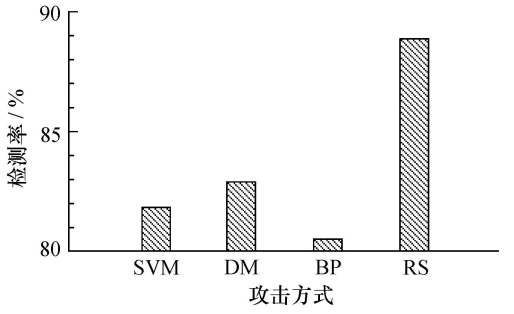

图 2 几种入侵检测方法的性能比较

通过对比的实验结果可以看出,粗糙集方法较其他入侵检测方法在检测率和错检率方面有了一些提高. 由于数据挖掘应用于入侵检测需要大量的数据支持,采用基于支持向量机和 BP神经网络的检测方法的训练速度和运算量都比较大等原因,使得基于 BP神经网络、支持向量机和数据挖掘的检测算法的有效性和检测率明显低于基于粗糙集的检测方法,这也充分证明了粗糙集理论应用于系统入侵检测的有效性.

5 结 论

本文从检测规则的角度思考问题,借鉴粗糙集思想,提出一种基于粗糙集的入侵检测算法,将粗糙集算法和入侵检测技术结合起来实现系统安全检测. 粗糙集在对不确定和不完备信息的处理上具有很大的优势,实验结果表明,粗糙集方法比其他检测方法在检测率和错检率上都有了较大的改进,可以提供比较准确的检测报告,从而能够为信息系统提供高效的入侵检测服务.

参考文献 (References)

[1] Bace R. Intrusion Detection [M]. New York: Macmillan Technical Publishing, 2000.

[2] 蒋建春, 马恒太, 任党恩, 等. 网络安全入侵检测: 研究综述 [J]. 软件学报, 2000, **11**(11): 1460 - 1466.
JIANG Jianchun, MA Hengtai, REN Dangen, et al. A survey of intrusion detection research on network security [J]. *Journal of Software*, 2000, **11**(11): 1460 - 1466. (in Chinese)

[3] 蔡忠闽, 管晓宏, 邵萍, 等. 基于粗糙集理论的入侵检测新方法. 计算机学报 [J], 2003, **26**(3): 361 - 366.
CAI Zhongmin, GUAN Xiaohong, SHAO Ping, et al. A new approach to intrusion detection based on rough set theory [J]. *Chinese Journal of Computers*, 2003, **26**(3): 361 - 366. (in Chinese)

[4] Forrest S, Perrelason A S, Allen L. Self-nonself discrimination in a computer [C]// Rushby J, Meadows C. Proceedings of the 1994 IEEE Symposium on Research in Security and Privacy. Oakland CA: IEEE Computer Society Press, 1994: 202 - 212.

[5] 饶鲜, 董春曦, 杨绍全. 基于支持向量机的入侵检测系统 [J]. 软件学报, 2003, **14**(4): 789 - 803.
RAO Xian, DONG Chunxi, YANG Shaoquan. An intrusion detection system based on support vector machine [J]. *Journal of Software*, 2003, **14**(4): 789 - 803. (in Chinese)

[6] Ghosh A K, Michael C, Schatz M. A real-time intrusion system based on learning program behavior [C]// Debar H, Wu S F (eds). Recent Advances in Intrusion Detection (RAID 2000). Toulouse: Spinger-Verlag, 2000: 93 - 109.

[7] 杨武, 云晓春, 李建华. 一种基于强化规则学习的高效入侵检测方法 [J]. 计算机研究与发展, 2006, **43**(7): 1252 - 1259.
YANG Wu, YUN Xiaochun, LI Jianhua. An efficient approach to intrusion detection based on boosting rule learning [J]. *Journal of Computer Research and Development*, 2006, **43**(7): 1252 - 1259. (in Chinese)

[8] Lee W, Stolfo S J. A data mining framework for building intrusion detection model [C]// Proceedings of the 1999 IEEE Symposium on Security and Privacy. Oakland, CA: IEEE Computer Society Press, 1999: 120 - 132.

[9] Pawalk Z. Rough sets [J]. *Int J Computer and Information Sci*, 1982, **11**(5): 341 - 356.

[10] Ziako W. Rough sets: Trends, challenges, and prospects [C]// Ziako W, Yao Y (eds). Rough Sets and Current Trends in Computing (RSCTC 2000). Banff: Springer-Verlag, 2001: 1 - 7.