

基于新的条件熵的入侵检测算法

罗晓¹, 于磊^{1,2}, 罗谦¹

(1. 中国民用航空局第二研究所, 四川 成都 610041;

2. 西南交通大学 信息科学与技术学院, 四川 成都 610031)

摘要:在分析了现有的入侵检测方法的基础上,为了降低入侵检测系统的错检率、降低漏检率和提高实时性,提出了一种新的检测方法:基于新的条件熵的入侵检测算法。本算法在考虑信息论有关理论的基础上,利用信息熵的知识对收集到的数据进行离散化。通过分析离散化后的数据,利用新的条件熵的知识约简方法去除冗余属性,生成检测规则,然后用来分析入侵数据。实验结果表明:基于新的条件熵的入侵检测算法与基于BP神经网络和支持向量机的入侵检测算法比较,可以有效地提高入侵检测系统的检测率,降低错检率。该算法的检测率提高7%左右,能为信息系统提供很好的入侵检测服务。

关键词:新的条件熵;离散化;入侵检测;知识约减

中图分类号:TP393.08

文献标识码:A

文章编号:1673-629X(2011)08-0237-04

Adaptive Intrusion Detection Algorithm Based on New Conditional Entropy

LUO Xiao¹, YU Lei^{1,2}, LUO Qian¹

(1. The Second Research Institute of CAAC, Chengdu 610041, China;

2. School of Information Science and Technology, Southwest Jiaotong University, Chengdu 610031, China)

Abstract:Based on the analysis of the current intrusion detection approaches, existing security detection systems have many problems such as wrong detection of intrusions, missed intrusions, poor real-time performance, bring up a new detection method, namely adaptive intrusion detection algorithm based on new conditional entropy. In considering the theories related to information theory, this algorithm firstly discrete the collected data use the knowledge of information entropy, then analyze the discrete data, remove the redundant attributes by reduction method related to conditional entropy knowledge, finally generate a new detection rules for the further analysis of intrusion data. The experimental result shows that is more efficient than algorithms based on BP neural networks and vector machines; thereby, this detection algorithm can effectively improve the intrusion detection system's detection rate, and reduce the error detection rate, and this detection algorithm can improve the detection ratio by about 7% and reduce the wrong detection ratio. The system provides detection service effective for information systems, as well.

Key words:new conditional entropy; discretization; intrusion detection; knowledge reduction

0 引言

入侵检测系统(Intrusion Detection System, IDS)是一套监控计算机系统或网络系统中发生的事件,根据规则进行安全审计的软件或硬件系统^[1,2]。由传统的电子数据处理技术、安全审计检查技术以及统计分析技术发展而来的入侵检测技术,在很多的安全系统中得到应用,它主要通过检查有关的审计数据,以判断系

统中是否有违背安全策略或计算机系统的安全行为。

现有的入侵检测技术已经得到了很大的发展,但是由于网络传输速度远远超过检测速度,有较高的误报率和漏报率产生。另外现有的入侵检测系统体系结构本身也存在许多问题,如其本身构建易受攻击等问题^[3]。尤其是近年来,网络攻击手段的多样化以及新型攻击方式的不断出现,加之安全检测技术的自身缺陷,给整个社会带来了巨大的损失。因此,提出一种能够检测潜在未知攻击的异常检测方法引起来各方面学者的很大兴趣,目前已经存在的异常检测算法和模型主要有:文献[4]提出了基于统计的入侵检测模型,但这种方式很难选取一个合理的阈值来判断一个模式是否是异常模式,阈值的选取不合理很容易导致系统

收稿日期:2011-01-20;修回日期:2011-04-01

基金项目:中国民用航空局科研项目(MHRD200924)

作者简介:罗晓(1970-),男,高级工程师,主要研究领域为机场信息集成技术、计算机仿真、数据库技术;于磊,硕士研究生,主要研究领域为粒计算、人工智能;罗谦,博士研究生,主要研究领域为数据挖掘、进化计算、企业智能计算。

的误报和漏报。文献[5]提出区分正常与非正常的数据是入侵检测技术的一个核心过程,在免疫模型的基础上提出一种新的检测技术;文献[6]利用神经网络来提取特征和分类;文献[7]从数据挖掘技术角度探讨了入侵检测的实现问题。文献[5~7]所提出的方法都需要大量或者是完备审计数据集才能达到比较合理的检测性能,而且这些方法都需要长时间的进行训练,在小样本的情况下,以上的方法就不太实用。文献[8]根据强化规则学习方法提出了一种检测技术,该技术将在入侵检测技术中使用规则学习算法,使系统的误报率有一定的降低;文献[9]把支持向量机技术应用于入侵检测系统,但是该方法同样训练时间较长,利用 SVM 对大规模数据进行训练时,需要占用很大的内存空间,甚至会因内存不够而无法训练,也就是利用该方法建立入侵检测模型较为困难;文献[10]从检测规则的角度,借鉴粗糙集的思想进行检测,在对不确定和不完备信息的处理上有很大的优势,但是由于数据收集后离散化不合理,打破了利用粗糙集作为算法核心的特殊性,等频划分决策表后得到离散化数据,这种离散化后的数据降低了决策表的相容性,使算法在规则泛化能力方面也大大降低。

另外,文献[11]考虑到经典的粗糙集理论决策表知识约简方法仍存在一定不足,提出了一种新的计算条件熵的方法,该方法在一定的条件下,处理知识约简有其独特的优点。文中也将借鉴该条件熵的提取来优化文中提出的算法。

将粗糙集理论和入侵检测技术相结合,利用决策表信息熵的角度的一种新的连续属性离散化算法,对由实际数据构成的决策表中各个属性值用离散值表达。在此基础上得到离散化后的数据,利用新的条件熵的知识约减方法进行处理,产生检测系统的分类检测规则,判断区分正常与非正常的数据行为,从而有效地降低系统的漏报率与错检率。

1 有关概念理解

文中只介绍利用粗糙集的离散化处理连续属性的几个基本概念,具体关于条件熵与信息熵的有关概念和性质,可参见文献[11,12]。

定义 1: 设决策表 (U, R, V, f) , $U = \{x_1, x_2, \dots, x_n\}$ 是论域, $R = C \cup \{d\}$, $V_d = \{1, 2, \dots, r(d)\}$ 为决策属性的值域,对每一个连续的条件属性 $a \in C$, 论域中的个属性值经过排序后可以得到 $l_a = v_0^a < v_1^a < \dots < v_{n_a}^a = r_a$, 候选断点可以定义为

$$c_i^a = (v_{i-1}^a + v_i^a) / 2 \quad (i = 1, 2, \dots, n_a) \quad (1)$$

定义 2: 另 $X \subseteq U$, $|X|$ 代表属于 U 的子集实例个

数, k_j 为决策属性是 $j (j = 1, 2, \dots, r(d))$ 的实例个数为, 定义此信息熵为

$$H(X) = - \sum_{j=1}^{r(d)} p_j \log_2 p_j, \quad p_j = \frac{k_j}{|X|} \quad (2)$$

另外说明: 信息熵 $H(X)$ 越小, 决策的混乱程度越小, 决策值越单一。

定义 3: 决策表的相容度。设 $j (j = 1, 2, \dots, r(d))$ 为断点 c_i^a 的决策属性值, 有关决策属性值所属的实例中, $l_j^X(c_i^a)$ 为属性 a 的值小于断点值 c_i^a 且属于集合 X 的实例个数, $r_j^X(c_i^a)$ 为大于断点 c_i^a 的实例个数。

$$\text{令 } l^X(c_i^a) = \sum_{j=1}^{r(d)} l_j^X(c_i^a) \quad (3)$$

$$r^X(c_i^a) = \sum_{j=1}^{r(d)} r_j^X(c_i^a) \quad (4)$$

c_i^a 将集合 X 分成两个子集合, 分别是 X_l 和 X_r , 且

$$H(X_l) = - \sum_{j=1}^{r(d)} p_j \log_2 p_j, \quad p_j = \frac{l_j^X(c_i^a)}{l^X(c_i^a)} \quad (5)$$

$$H(X_r) = - \sum_{j=1}^{r(d)} q_j \log_2 q_j, \quad q_j = \frac{r_j^X(c_i^a)}{r^X(c_i^a)} \quad (6)$$

所以, 断点 c_i^a 对于集合 X 的信息熵为

$$H^X(c_i^a) = \frac{|X_l|}{|U|} H(X_l) + \frac{|X_r|}{|U|} H(X_r) \quad (7)$$

定义 4: 假设 $L = \{Y_1, Y_2, \dots, Y_m\}$ 为决策表已获得断点集合, 在断点 P 的划分下得到等价类 L , 设另外一个断点 c 且 $c \notin P$, 将 c 加入后, 得到

$$H(c, L) = H^{Y_1}(c) + H^{Y_2}(c) + \dots + H^{Y_m}(c) \quad (8)$$

依据相关定义得知, $H(c, L)$ 越小表明断点 c 的加入使得决策属性值越精确, 也更加单一, 因此 $H(c, L)$ 为断点 c 对整个决策的重要性的体现。

2 基于新的信息熵的入侵检测模型

文中根据新的条件熵的知识, 结合已有的入侵检测模型提出了一种新的入侵检测模型, 如图 1 所示, 先对网络数据进行收集, 然后进行数据规整, 利用处理好的数据, 对连续性的条件属性进行基于粗糙集的离散化处理, 利用新的条件熵的知识进行属性约简, 约减连续属性后产生规则, 构造该检测系统的规则库, 生成入侵检测器, 系统利用初次建立起来的入侵检测模型进行检测, 并在以后的逐步运行过程中改进和完善模型。

根据以上的检测模型可以得知, 文中构建的监测系统主要有以下几个问题:

一个入侵检测系统对数据的收集是必不可少的, 数据收集阶段作为该模型的第一个阶段, 一个入侵检测系统要面临大量的用户入侵数据、网络入侵数据以及用户入侵行为的行为数据。数据收集就是要把收集到的数据进行规整和预处理, 网络数据包的分析是整

个数据处理的难点和侧重点,网络数据更加具有复杂多样性,为了提高处理效果,需对收集到的网络数据进行一些简单的基本的处理。

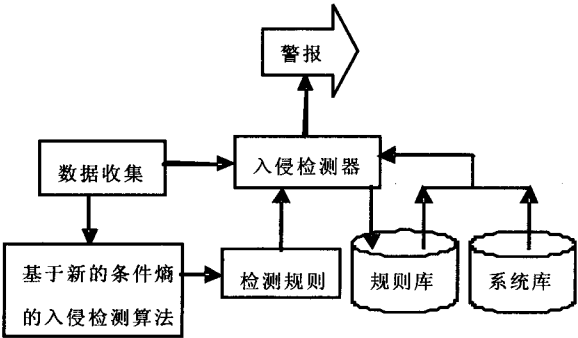


图 1 基于信息熵的入侵检测模型

基于新的条件熵的入侵检测算法,是整个入侵检测系统的核心。文中所提出的入侵检测算法主要是由两部分组成,首先是对收集到的数据进行离散化处理,然后对入侵检测属性进行约简,去除冗余入侵属性,产生决策规则。利用粗糙集的有关理论和方法进行规则的约简,能够很好地挖掘出潜在的、有效的规则,生成较好的规则库和样本集,能够使系统的检测能力进一步提高,所以本算法同文献[11]一样,也使用同样的约减方法。在利用粗糙集的相关理论和方法对数据进行分析,要对决策表中各个数据的属性值须使用离散值表达,这样能够更好地获取知识。但是对粗糙集连续属性离散化的方法一般是采用其他领域已经存在的方法,如文献[10]采用等频划分的方法,这些方法就没有考虑到粗糙集理论的他属性,离散化后的数据相关性降低,破坏原有数据的相容性,使以后知识约减提取检测规则的能力下降,所以文中将采取基于信息熵的离散化方法进行离散化,同时也会采取有关新的条件熵的属性约简方法进行决策属性约简,产生决策规则。

利用粗糙集的理论对数据约简后,将相关数据的多余属性值或者是无关属性值进行删除,对决策表进行精简和构建,利用决策表导出初始检测规则。对形成的初始检测规则进行检测和核实(利用入侵检测器),放到规则库,结合系统库构成新的安全检测器,这样随着系统的不断运行,该系统的入侵能力也不断地加强,也就是说该入侵检测系统具有了自适应能力。

3 基于新的条件熵的入侵检测算法

整个入侵检测系统分为三部分:数据收集、利用基于新的条件熵的入侵检测算法生成检测规则和检测模型的更新三个阶段,阶段二中生成检测规则的方法使用文献[11]提供的方法,其思想可以如下描述:

阶段一 对收集到的数据进行离散化。

(说明: H = 决策表信息熵; P = 已选取的断点集合; B = 待选断点集合; L = 实例按断点集合被划分成的等价类集合。)

- 步骤 1 $P = \Phi; L = \{U\}; H = H(U);$
 - 步骤 2 对每一个 $c \in B$, 计算 $H(c, L);$
 - 步骤 3 若 $H \leq \min\{H(c, L)\}$, 则结束;
 - 步骤 4 选择使 $H(c, L)$ 最小的断点 c_{\min} 加到 P 中; $H = H(c, L) \quad B = B - \{c\};$
 - 步骤 5 设 $M \in L$, 若把等价类 M 划分成为 $M1$ 和 $M2$, 那么就从 L 中去掉 M , 让后把等价类 $M1$ 和 $M2$ 加到 L 中;
 - 步骤 6 如果 L 中各个等价类中的实例都具有相同的决策, 则结束; 否则转到步骤 2。
- 阶段二 设置入侵检测规则。
- 步骤 7 利用文献[11]介绍的新的信息熵的属性约减方法进行属性约简, 输出一个最小相对属性约简, 导出检测规则, 结束。
- 阶段三 进入入侵检测分析。
- 步骤 8 安全检测器将依据约减得到的入侵检测规则, 对数据进行测试和分析, 如果测试或分析合理则转为 $K = K + R$ 。
 - 步骤 9 对检测器选取不同的训练样本进行多次训练, 利用测试结果完善安全检测模型, 更新检测器, 直到满足一定的误用率、虚警率和漏检率为止。
 - 步骤 10 算法结束。

4 算法的评价和实验分析

将入侵数据收集一起并进行测验, 将连续的入侵检测数据进行离散化处理, 得到测试数据和训练数据分别是 4825 条、6875 条, 进行基于新的条件熵的入侵检测算法处理, 进行多次测试, 得到实验数据见表 1。

表 1 检测率和错检率的测试数据

攻击方式	检测率(%)	错检率(%)
正常数据	93.87	4.02
U2R 攻击	76.83	18.13
DOS 攻击	84.98	11.28
R2L 攻击	82.12	14.37
Probe 攻击	81.69	8.56

实验结果可以得出, 对大量的网络入侵数据利用新的条件熵的入侵检测方法进行检测, 算法在漏检率和错检率上得到了改进。但由于该算法使用知识约减的方法, 对属性进行约减, 使得数据的连续性受到一定的影响, 会出现个别信息丢失的现象, 但是该现象是使用知识约减的方法不可避免的, 导致个别数据的检测率偏低而错检率较高。但是与其他的攻击方式的效果来比较, 该算法的效率和有效性得到了一定的提高。

基于信息熵的入侵检测算法与其他入侵检测方法相比,得到检测率数据见表 2。

通过表 2 可以看出,基于新的条件熵的方法在漏检率和错检率方面得到了改善。由于基于数据挖掘(DM)的方法需要大量的数据支持,在小样本数据中该方法的优越性无法体现。支持向量机(SVM)和 BP 神经网络的检测方法运算量较大且复杂,而且要求较高的训练速度。所以 SVM、DM 和 BP 的方法在检测率上不如文中的方法。

表 2 各检测方法错检率比较

检测方法	检测率(%)
SVM	81.75
DM	83.02
BP	80.32
文中算法	84.96

5 结束语

从检测规则的角度思考问题,借鉴信息论中的有关思想,提出了一种基于新的条件熵的入侵检测算法,将粗糙集相关的知识约简算法,以及利用粗糙集技术的新的离散化算法和入侵检测技术结合起来构建一个完善的入侵检测体系。有关粗糙集的相关理论和方法善于对不确定和不完备信息的处理,使用信息熵的有关知识对收集到的数据进行处理,更有利于知识约减和检测规则的提取。实验结果表明,基于新的条件熵的入侵检测算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,有较高实用性。

参考文献:

[1] Bace R. Intrusion Detection[M]. New York: Macmillan Tech-

(上接第 219 页)

直观的认识。

参考文献:

- [1] 吴志忠. 移动通信无线电波传播[M]. 北京: 人民邮电出版社, 2002.
- [2] 佟学俭, 罗 涛. OFDM 移动通信技术原理与应用[M]. 北京: 人民邮电出版社, 2003: 45-60.
- [3] 刘 然, 江修富, 郝建华. 利用训练符号进行一系统粗初定时同步算法的研究[J]. 国外电子测量技术, 2008, 27(7): 37-39.
- [4] 庄明洁, 郭东辉. 移动通信中无线信道特性的研究[J]. 电讯研究, 2004, 44(5): 18-21.
- [5] 李向宁, 谈振辉. OFDM 基本原理及其在移动通信中的应用[J]. 重庆邮电学院学报, 2003(6): 55-59.
- [6] 朱近康. 无线信道的应用模型和估计[J]. 中兴通讯技术,

nical Publishing, 2000.

- [2] 肖竞华, 卢 娜. 基于网络的入侵检测系统的研究及实现[J]. 计算机技术与发展, 2007, 17(2): 242-244.
- [3] 蔡忠闽, 管晓宏. 基于粗糙集理论的入侵检测新方法[J]. 计算机学报, 2003, 26(3): 361-366.
- [4] Debar H, Becker M, Siboni D. A neural network component for an intrusion detection system[C]// In: Proc. 1992 IEEE Symposium on Security and Privacy. Los Alamitos: IEEE Computer Society Press, 1992: 240-251.
- [5] Forrest S, Perrelason A S, Allen L. Self-nonsel self 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.
- [6] Ghosh A K, Michael C, Schatz M. A real-time intrusion system based on learning program behavior [C]// Debar H, Wu S F. Recent Advances in Intrusion Detection (RAID 2000). Toulouse: Springer-Verlag, 2000: 93-109.
- [7] 程玉青, 梅登华, 陈龙飞. 基于数据挖掘的入侵检测系统模型[J]. 计算机技术与发展, 2009, 19(12): 123-126.
- [8] 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.
- [9] 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.
- [10] 赵曦滨, 井然哲, 顾 明. 基于粗糙集的自适应入侵检测算法[J]. 清华大学学报, 2008, 48(7): 1165-1168.
- [11] 徐久成, 孙 林, 马媛媛. 基于新的条件熵的决策表约简方法[J]. 计算机工程与设计, 2008, 29(9): 2313-2316.
- [12] 翟俊海, 王熙熙, 张素芳. 信息粒度、信息熵与决策树[J]. 计算机工程与应用, 2009, 45(12): 126-128.

2003(10): 23-26.

- [7] 丁玉美, 高西全. 数字信号处理[M]. 西安: 西安电子科技大学出版社, 2001: 97-120.
- [8] Householder A S. The Theory of Matrix in Numerical Analysis [M]. New York: Dover Publications, 1964.
- [9] Bingham J A C. Multicarrier modulation for data transmission: an idea whose time has come[J]. IEEE Communication Magazine, 1990, 28(5): 5-14.
- [10] IEEE P80216-2004. Draft IEEE Standard for Local and Metropolitan Area Networks Part16: Air Interface for Fixed Broadband Wireless Access Systems[S]. [s.l.]: IEEE, 2004.
- [11] Air Interface for Fixed Broadband Wireless Access Systems [S]. IEEE Std. 802. 16d, 2004.
- [12] Nichols S J V. Achieving Wireless Broadband with WiMAX [J]. IEEE Comp, 2004, 37(6): 10-13.