文章编号: 1007-5321(2003)04-0085-05

基于用户击键数据的异常入侵检测模型

罗守山[1], 陈亚娟[2], 宋传恒[2], 王自亮[2], 钮心忻[2], 杨义先[2]
(1. 北京邮电大学 软件学院, 北京 100876; 2. 北京邮电大学 信息工程学院, 北京 100876)

摘要: 改进了 Apriori 关联规则算法, 通过对正常用户击键数据的规则挖掘, 建立用户的正常特征轮廓, 并以此对新用户的击键数据实行异常入侵检测. 实验结果表明, 本文提出的模型具有一定的入侵检测功能.

关 键 词: 网络安全; 数据挖掘; 关联规则; 入侵检测; 异常入侵检测
中图分类号: TN 393.08　　**文献标识码**: A

An Abnormal Intrusion Detection Model Based on User's Keystrokes

LUO Shou-shan[1], CHEN Ya-juan[2], SONG Chuan-heng[2],
WANG Zi-liang[2], NIU Xin-xin[2], YANG Yi-xian[2]
(1. Software Engineering School, Beijing University of Posts and Telecommunications, Beijing 100876, China;
2. Information Engineering School, Beijing University of Posts and Telecommunications, Beijing 100876, China)

Abstract: The Apriori association rules are improved. Through mining the rules of normal user's keystrokes, we set up the user's characteristic profiles. Then we use the profiles to do abnormal intrusion detection for a new user's keystroke data. The results show that this model can realize some functions of intrusion detection.

Key words: network security; data mining; association rules; intrusion detection; abnormal intrusion detection

入侵检测系统的研究是近年来网络安全领域的一个研究热点, 而入侵检测算法是实现该系统的核心. 已经有一些学者提出了基于不同知识背景的入侵检测算法. 本文利用数据挖掘技术, 通过挖掘用户的键盘行为特征, 建立相关的数学模型, 实现系统入侵检测.

1 异常入侵检测

入侵可以定义为[1]: 任何企图破坏计算机资源的完整性、机密性和可用性的行为的集合. 入侵检测就是检测入侵活动, 并采取对抗措施. 传统的入侵检测使用的方法主要分为两类:

收稿日期: 2002-08-05
基金项目: 国家重点基金研究发展规划项目资助(G1999035805); 国家自然科学基金项目资助(60073049); 国家杰出青年基金项目资助(69425001); 高等学校骨干教师资助计划
作者简介: 罗守山(1962—), 男, 北京邮电大学教授. E-mail: buptlou@263.net

异常检测与滥用检测.异常入侵检测的思想是Anderson于1980年首先提出[2], 1987年,Denning给出了一个异常入侵检测系统的模型[3].异常入侵检测的主要前提条件是将入侵性活动作为异常活动的子集,理想状况是异常活动集与入侵性活动集等同,这样,若能检测所有的异常活动,则可检测所有的入侵性活动.但是,入侵性活动并不总是与异常活动相符合.它存在4种可能性[4]: (1)入侵性而非异常; (2)非入侵性且异常; (3)非入侵性且非异常; (4)入侵且异常.目前,常用的异常入侵检测方法见文献[5, 6]

2 利用关联规则实现数据挖掘

近年来,数据挖掘技术的研究引起了国际人工智能和数据库等领域的专家学者的广泛关注.关联规则是由R. Agrawal等人首先提出[7].该规则挖掘在商业等领域的成功应用,使它成为数据挖掘中最成熟、最重要、最活跃的研究内容[8~11].

关联规则的挖掘问题就是在事务数据库D中找出具有用户指定的最小支持度minsup和最小置信度minconf的关联规则.它可以分解为以下2个子问题:

(1) 找出事务数据库D中所有具有用户指定的最小支持度minsup的大项集.

(2) 利用大项集产生关联规则.对于任意的大项集A和A的任何非空子集B,如果Support(A)/Support(B)≥minconf,则生成关联规则$R: B \Rightarrow (A - B)$.其支持度和置信度分别为:

$$\text{Support}(R) = \text{Support}(A \cup B) = \text{Support}(A),$$

$$\text{Confidence}(R) = \text{Support}(A \cup B)/\text{Support}(B) = \text{Support}(A)/\text{Support}(B)$$

其中,第二个问题比较简单.目前大多数研究集中在第一个问题上.

3 利用Apriori关联规则实现基于键盘行为模式的入侵检测

3.1 Apriori关联算法

首先,对Apriori关联算法做一个介绍.

Apriori算法[12].算法流程如下(为了将Apriori算法用于入侵检测,本文对其做了一些修改,符号·之后的部分是作者增加的内容):

1) $L_1 = \{\text{large1-itemsets}\}$; ·$M_1 = L_1$;

2) for($k = 2; L_{k-1} \neq \varnothing; k++$) do begin

3) C_k = apriori-gen(L_{k-1});

4) forall transactions $t \in D$ do begin

5) C_t = subset(C_k, t);

6) forall candidates $c \in C_t$ do

7) c.count++; ·if Apriori-con(c), then c.realcount++ end if

8) end

9) $L_k = (c \in C_k \mid c.\text{count} \geq \text{min sup})$ ·$M_k = (c \in C_k \mid c.\text{realcount} \geq \text{min sup})$

10) end

11) Answer = $\cup_k L_k$; ·Answer = $\cup_k M_k$

函数apriori-gen()的作用是产生候选大k项集C_k.该函数分为合并,剪除2个部分.在合并部分,通过合并2个$k-1$大项集L_{k-1p}, L_{k-1q}得到一个可能的k大项集.算法如下:

insert into C_k

select $p.item_1$, $p.item_2$, …, $p.item_{k-1}$, $q.item_{k-1}$ from $L_{k-1}p$, $L_{k-1}q$, where $p.item_1 = q.item_1$, … $p.item_{k-2} = q.item_{k-2}$, $p.item_{k-1} < q.item_{k-1}$;

在剪除部分, 对于 C_k 中的候选 k 大项集 c, 如果 C_k 的 $(k-1)$ 维子集不在 L_{k-1} 中, 就将其剪除. 算法如下:

```
forall itemset c∈C_k do
    forall (k-1) subsets s of c do
      if (s∉L_{k-1}) then
        delete c from C_k.
```

其次, 本文对 Apriori 算法做了修改. 原因是想让挖掘出来的关联规则适合于入侵检测.

这里考虑的数据是用户的键盘数据, 所以, 该问题的背景应该考虑由该数据产生的项目集中的项目的连续次序. 比如说, 对于用户多次键入 tion, 期望能够挖掘出如 $t \& i \Rightarrow o$; $t \Rightarrow i \& o$; $i \& o \Rightarrow n$; $i \Rightarrow o \& n$ 的 3 元素关联规则, 而不期望挖掘出形如: $t \& o \Rightarrow n$ 和 $t \Rightarrow o \& n$ 形式的规则. 为此, 本文增加了函数 Apriori-con(c)

```
Apriori-con(c)
{
if c 连续 then return 1;
        else return o
end if
}
```

3.2　挖掘用户正常的键盘工作模式及其实现入侵检测

挖掘用户的正常工作模式原理如下:

第一步: 收集数据. 统计用户在一段时间内的键盘、鼠标的工作数据, 并将这些数据分为两组, A 组用于挖掘规则, B 用于检验.

第二步: 挖掘关联规则. 对于给定的支持度和置信度, 利用 Apriori 算法挖掘 A 组数据中的关联规则集 $A1$.

第三步: 检验关联规则. 对 B 组数据进行同样支持度、置信度下的挖掘, 得关联规则集 $B1$. 比较规则 $A1$ 与 $B1$, 规定一个满意度(0 与 1 之间的一个实数, 通常应大于 0.7), 如果 2 次挖掘生成的规则集的交集在 $A1$、$B1$ 中所占的百分比均大于满意度, 则认为挖掘规则成功, 即规则能够代表用户正常的工作模式, 规则集 $A1$ 为该用户的轮廓特征. 否则, 认为现有的数据不能代表用户的工作模式, 应该继续搜集用户数据.

第四步: 利用规则进行入侵检测. 首先约定一个检测阈值(0 与 1 之间的一个实数, 通常应大于 0.7). 统计新用户的键盘工作数据, 使用相同的预处理、利用相同的支持度、置信度挖掘关联规则, 并跟正常用户的轮廓特征做比较, 如果大于检测阈值, 则接受该用户, 否则, 启动入侵检测报警装置. 检测阈值的定义如下:

定义 1　检测阈值 $= \dfrac{\text{新用户挖掘出的规则在用户轮廓特征库中所占的数量}}{\text{新用户数据中挖掘出的规则数量}}$.

这里, 对新用户数据的挖掘可以采用定时间隔的方式.

在实验中, 统计了一个正常用户的键盘工作方式, 经过预处理, 得到了一个由 6 316 个单词、数字组成的集合, 即生成了一个有 6 316 个项目组成的事务数据库, 每一个项目是一个单

词或数字．用支持度、置信度分别为0.01、0.01，挖掘长度为5，在Apriori算法下，挖掘出关联规则168个，具体分布如表1所示．

表1 Apriori规则挖掘结果

	2元素关联	3元素关联	4元素关联	5元素关联
正常用户(6 316项)	109	52	7	0
新用户1(14 039项)	107	26	4	0
新用户2(1 540项)	153	105	19	2

其含义如下：2元素关联是指形如$a \Rightarrow b$(支持度、置信度)的规则．3元素关联是指形如$a \Rightarrow b \ \& \ c$(支持度、置信度)，或$a \ \& \ b \Rightarrow c$(支持度、置信度)等．如对于正常用户，挖掘出形如$a \Rightarrow b$(0.022 422 2, 0.094 778 9)，$r \Rightarrow i$(0.030 001 6, 0.503 989)等共109个；形如$s \ \& \ t \Rightarrow r$(0.010 105 8, 0.361 318)的规则52个，形如$t \ \& \ i \ \& \ o \Rightarrow n$(0.036 633 5, 0.743 707)的规则7个．且如上规则通过满意度0.8的测试．

随后，以正常用户的轮廓特征为基础，进行规则挖掘(如表1)，并利用2个新用户的规则，进行阈值检测如表2所示．

表2 阈值检测

	2元素满意度	3元素满意度	4元素满意度	5元素满意度	结论
新用户1	$\frac{98}{107}=0.916$	$\frac{21}{26}=0.808$	$\frac{4}{4}=1.000$	$\frac{0}{0}$	正常
新用户2	$\frac{63}{153}=0.412$	$\frac{18}{105}=0.171$	$\frac{2}{19}=0.105$	$\frac{0}{2}=0.000$	不正常

表2的含义如下：以新用户1为例，在107个二元素关联中，有98个出现在正常用户的二元素关联集中，故可计算出其检测阈值为$\frac{98}{107}=0.916$．约定检测阈值为0.8，结果表明，系统认为新用户1是正常用户，而新用户2为入侵者．

3.3 对模型的评价

值得注意的是，提取用户键盘工作行为特征的入侵检测模型是基于如下假设：每一个用户的工作内容存在着一个轮廓特征，并且该用户在检测阶段的工作内容与其轮廓特征没有显著的变化．本文设计的检测模型可以用来对正常用户的特征轮廓进行动态更新．当系统确认用户身份后，可以将挖掘出的规则加入到正常用户的规则库中．经过对多组用户数据的规则挖掘，并分析检测结果，我们得出如下经验：置信度的选择一般在0.3左右为佳．

4 结 论

利用数据挖掘技术实现入侵检测在国内外已经有学者对此进行过研究，而利用数据挖掘技术，通过挖掘用户的键盘行为特征实现入侵检测，还没有见到过相关的研究报告．本文在此方面做了探索，首先改进了Apriori关联规则算法，并利用它实现对用户键盘输入数据的挖掘，建立正常用户的键盘工作模式，实现异常入侵检测．实验结果表明，对于盗用合法用户口令的非法用户，以及合法用户的越权操作，本文提出的模型能够实现一定的入侵检测功能．

参考文献:

[1] Heady R, Luger G, Maccabe A, et al. The architecture of a network level intrusion detection system [R]. Technical Report CS90_2, University of New Mexico, Department of Coomputer Science, 1990.

[2] Anderson J P. Computer security threat monitoring and surveillance[R]. Technical Report, James P Anderson, Co., Fort Washington, Pennsylvania, 1980.

[3] Denning D E. An intrusion-detection model[J]. IEEE Trans on Software Eng, 1987, 13 (2): 222-232.

[4] 蒋建春, 马恒太, 任党恩, 卿斯汉. 网络安全入侵检测: 研究综述[J]. 软件学报, 2000, 11(11): 1 460-1 466.

[5] 蒋建春, 冯登国. 网络入侵检测原理与技术[M]. 北京: 国防工业出版社, 2001. 29-41.

[6] Rebecca Gurley Bace 著, 陈明奇 等译. 入侵检测[M]. 北京: 人民邮电出版社, 2001. 100-101.

[7] Agrawal R, Imielinski T, Swami A. Mining association rules between sets of items in large databases [A]. Proceedings of the ACM SIGMOD Conference on Management of Data[C]. Washington D. C, 1993, 5.

[8] M S, et al. Data mining: an overview from database perspective[J]. IEEE Transactions on Knowledge and Data Engineering, 1996, 8(6): 866-883.

[9] 铁治欣, 陈奇, 俞瑞钊. 关联规则采掘综述[J]. 计算机应用研究, 2000, 10(1): 1-5.

[10] 赵亮, 王培康. 关联规则发现综述[J]. 计算过程与应用, 2001, 13(18): 94-98.

[11] 朱绍文 等. 关联规则挖掘技术及发展动向[J]. 计算机工程, 2000, 26(9): 4-6.

[12] Agrawal R, Srikant R. Fast algorithms for mining association rules[A]. In Proceedings of the 20th International Conference on Very Large Data Bases[C]. Santiago: Chile, 1994, 487-499.

(上接第 69 页)

参考文献:

[1] Austin M D, Stuber G L. In service signal quality estimation for TDMA cellular systems[C]. Proc PIMRC, 1995. 836-840.

[2] Andersin M, Mandayam N B, Yates R D. Subspace based estimation of the signal to interference ratio for TDMA cellular systems[C]. Proc VTC, Atlanta, GA, 1996. 1 155-1 159.

[3] Turkboylari M, Stuber G L. An efficient algorithm for estimating the signal-to-interference ratio in TDMA cellular systems[J]. IEEE Trans Commun, 1998, 46(6): 728-731.

[4] RECOMMDATION ITU-R M. 1225- 2000, Guidelines for evalulation of radio transmission technologies for imt[S].