一种改进的 IDS 异常检测模型

孙宏伟[1),2)] 田新广[1),2)] 李学春[2)] 张尔扬[1)]

[1)](国防科技大学电子科学与工程学院 长沙 410073)

[2)](北京首信集团研究院 北京 100016)

摘 要 基于机器学习的异常检测是目前 IDS 研究的一个重要方向. 该文对一种基于机器学习的用户行为异常检测模型进行了描述,在此基础上提出一种改进的检测模型. 该模型利用多种长度不同的 shell 命令序列表示用户行为模式,建立多个样本序列库来描述合法用户的行为轮廓,并在检测中采用了以 shell 命令为单位进行相似度赋值的方法. 文中对两种模型的特点和性能做了对比分析,并介绍了利用 UNIX 用户 shell 命令数据进行的实验. 实验结果表明,在虚警概率相同的情况下改进的模型具有更高的检测概率.

关键词 IDS;机器学习;异常检测;相似度

中图法分类号 TP18

An Improved Anomaly Detection Model for IDS

SUN Hong-Wei[1),2)] TIAN Xin-Guang[1),2)] LI Xue-Chun[2)] ZHANG Er-Yang[1)]

[1)](*School of Electronic Science and Engineering, National University of Defense Technology, Changsha* 410073)

[2)](*Research Institute of Beijing Capitel Corporation, Beijing* 100016)

Abstract The application of machine learning technique to anomaly detection acts as one of the important directions of research on IDS. This paper introduces an user behavior anomaly detection model based on machine learning originated mainly by Terran Lane, and presents an improved anomaly detection model. It uses shell command sequences of variable lengths to represent user behavior patterns and construct more than one libraries of command sequences to represent normal user behavior profiles. While performing detection, the model mines behavior patterns in the stream of shell command sequences generated by the current user, and evaluates the similarity for each shell command according to the length of the sequence, i. e. the behavior pattern which it belongs to. The similarities of the commands are then filtered and act as the measure to determine whether the behavior of the current user is normal or not. The performance of the model is tested by computer simulation with UNIX users' shell command data. The results show it has higher detection accuracy than Terran Lane's model.

Keywords IDS; machine learning; anomaly detection; similarity

1 IDS 的两类检测技术及研究现状

入侵检测技术主要有两种类型,即误用检测和异常检测. 误用检测的核心是建立并维护一个入侵模式库,其特点是能够准确检测出已知的入侵,并报告出其类型,但是,模式库需要不断更新,而且难以检测未知的入侵. 异常检测则是建立系统或用户的

收稿日期:2002-06-20;修改稿收到日期:2003-07-11. 本课题得到中国普天首信集团重大科研项目(20011025)资助. **孙宏伟**,男,1964 年生,博士研究生,研究方向为网络安全、入侵检测. E-mail:sunhw@cpit.com.cn. **田新广**,男,1976 年生,博士研究生,研究方向为信号与信息处理、网络安全. **李学春**,男,1963 年生,教授,研究方向为网络安全. **张尔扬**,男,1941 年生,教授,博士生导师,主要研究方向为数字信号处理、通信网络.

正常行为模式库,通过被监测系统或用户的实际行为模式与正常行为模式之间的比较和匹配来检测入侵,其优点是不需要过多有关系统缺陷的知识,具有较强的适应性,能够实现对未知入侵的检测,但它存在虚警率高的缺点.目前,基于专家系统、模式匹配、概率统计等技术的误用检测模型已广泛应用于各类入侵检测系统(IDS).

由于入侵方法越来越多样化和综合化,单靠传统的检测技术已难以满足要求,基于智能技术的异常检测正成为 IDS 研究的一个重要方向.国内外已经开展了机器学习、神经网络、遗传算法等智能技术在异常检测中的应用研究,研究目标主要是提高检测系统的准确性、实时性、高效性以及自适应性,其中一些研究成果在检测性能和可操作性上已经接近或达到了实用化水平.本文介绍了 Terran Lane 等人提出的基于机器学习的定长命令序列异常检测模型,在其基础上提出一种变长命令序列检测模型,并介绍了利用 UNIX 用户 shell 命令数据进行了实验.实验结果证明了此模型在检测性能上的优势.

2　一种基于机器学习的用户行为异常检测模型

2.1　机器学习原理

机器学习是根据生理学、认知科学对人类学习机理的了解,借助机器(计算机系统)建立人类学习的计算模型和认知模型,发展各种学习理论和学习方法,研究通用的学习算法,在此基础上构建具有特定应用的面向任务的学习系统.图 1 给出了机器学习系统的通用模型.

图 1 中,环境为学习单元提供外界信息,学习单元利用该信息来建立知识库并对其做出改进(增加新知识或重新组织已有知识),执行单元利用知识库中的知识执行任务,任务执行后的信息又反馈给学习单元作为进一步学习的输入.

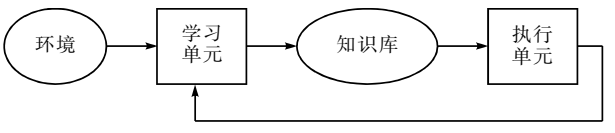

图 1　机器学习系统模型

机器学习研究的很大一部分工作集中在分类和问题求解这两个领域.学习单元是机器学习系统内实现学习方法功能的核心,它涉及处理外界信息的方式以及获取新知识过程中所用的推理方法.知识库用来存储知识,包括系统原有的领域知识(这种知识是长期的、相对稳定不变的)以及通过学习而获得的各种新知识(这种知识是短期的、变化的),选择何种知识表示对学习系统的设计起着非常重要的作用.执行单元是使得学习系统具有实际用途,同时又能够评价学习方法好坏的关键部分.机器学习目前已经发展了很多学习方法,例如实例学习、归纳学习、类比学习等,但是,这些方法均有其局限性.结合具体的应用领域,探讨新的学习方法和算法是目前的研究主流.

2.2　基于机器学习的定长命令序列检测模型

普渡大学的 Lane 等人将机器学习技术应用于异常检测,提出一种定长命令序列检测模型,其研究基于 UNIX 用户 shell 命令数据,以实验为主.该模型利用长度固定的 shell 命令序列表示用户的行为模式,建立一个样本序列库来描述一个(或一组)合法用户的正常行为轮廓,并定义命令序列之间的相似度;工作时,将被监测用户的命令序列同合法用户的命令序列样本库进行对比,根据两者的相似度来判别被监测用户行为的正常与否.模型具体描述如下:

(1)定义长度固定的命令序列,用命令序列表示用户的行为模式,建立一个样本序列库来描述一个(或一组)合法用户的正常行为轮廓.

设一个合法用户的正常训练数据为 $R=\{s_1,s_2,\cdots,s_r\}$,它是该用户在正常操作时所执行的长度为 r 的 shell 命令流,其中 s_i 表示按时间顺序排列的第 i 个 shell 命令.定义命令序列 $Seq_i=(s_i,s_{i+1},\cdots,s_{i+l-1})$,其中 l 为序列长度.相应的 shell 命令序列流可表示为 $S=\{Seq_1,Seq_2,\cdots,Seq_{r-l+1}\}$,其长度为 $r-l+1$.运用实例学习方法从 S 中选取若干个序列可建立样本序列库 L,它用于代表该合法用户的正常行为轮廓(当有多个合法用户时,可以针对每个用户建立一个样本序列库,然后将这些样本序列库组合在一起构成一个总的序列库,用它来代表这些合法用户整体的正常行为轮廓).选取样本序列的方法包括聚类、按出现概率提取、按时间顺序截取、随机选择等,这些方法直接关系到 L 对合法用户行为轮廓的反映程度,因而对模型的性能有很大影响.

(2)定义两序列之间的相似度函数,它用于表示两个序列所代表的行为模式之间的相似程度(相似

度函数值越大,表示相似程度也越大).在此基础上,定义一个序列同样本序列库的相似度函数,它用于表示此序列所代表的行为模式同合法用户正常行为轮廓的相似程度.

设两序列 Seq_a 和 Seq_b 的相似度函数为 $\mathrm{Sim}(Seq_a, Seq_b)$,可对它做多种定义,模型中定义了四种相似度函数,其中有两种函数不关心相邻命令之间的相关性,其余两种则考虑了这种相关性;函数的最大值随 l 的增大分别按多项式增长和指数增长.关心相邻命令之间的相关性且最大值按多项式增长的相似度函数定义如下[1]:

① 设 $c:=1, SIM:=0, j:=1$.

② 如果 $Seq_a(j)=Seq_b(j)$,则 $SIM:=SIM+c, c:=c+1$;否则,$SIM:=SIM, c:=1$.($Seq_a(j)$ 和 $Seq_b(j)$ 分别表示 Seq_a 和 Seq_b 中的第 j 个 shell 命令).

③ $j:=j+1$. 如果 $j\leq l$,返回执行步②;否则,$\mathrm{Sim}(Seq_a, Seq_b):=SIM$.

根据以上定义,当 $Seq_a(j)=Seq_b(j), 1\leq j\leq l$ 时(即两序列相同时),$\mathrm{Sim}(Seq_a, Seq_b)=l(l+1)/2$.

模型中序列 Seq_a 和样本序列库 L 的相似度函数定义为

$$\mathrm{Sim}(Seq_a, L)=\max_{Seq_i\in L}\{\mathrm{Sim}(Seq_a, Seq_i)\} \quad (1)$$

(3)模型工作时,对于被监测用户的 shell 命令序列流中的每个序列,计算它同样本序列库的相似度函数值,然后,将各个序列对应的相似度函数值进行加窗滤噪处理,得到最终的相似度输出值,对此值设定一个门限 λ,若它大于 λ,将被监测用户的当前行为判为正常行为,否则,将其判为异常行为.

设被监测用户在被监测时间内所执行的 shell 命令流为 $\bar{R}=\{\bar{s}_1, \bar{s}_2, \cdots, \bar{s}_r\}$,相应的 shell 命令序列流为 $\bar{S}=\{\overline{Seq}_1, \overline{Seq}_2, \cdots, \overline{Seq}_{r-l+1}\}$,其中 $\overline{Seq}_i=(\bar{s}_i, \bar{s}_{i+1}, \cdots, \bar{s}_{i+l-1})$.模型中有两种加窗滤噪算法可选择,分别是均值滤噪和中值滤噪,采用均值滤噪算法时的相似度输出值为

$$M(i, L)=\frac{1}{w}\sum_{j=i-w+1}^{i}\mathrm{Sim}(\overline{Seq}_j, L) \quad (2)$$

式中 $M(i,L)$ 表示被监测用户序列流 $\bar{S}$ 中第 i 个序列对应时刻的相似度输出值,w 表示窗长度;i 的初值为 w,其增长步长为 1.由于每个序列和每个命令(不包括 $\bar{R}$ 中的最后 $l-1$ 个命令)是一一对应的,所以 $M(i,L)$ 也是被监测用户命令流 $\bar{R}$ 中第 i 个 shell 命令对应时刻的相似度输出值.

以上模型中,有以下几个关键问题:序列长度的选择;样本序列的提取;相似度函数的定义;滤噪算法的选择.Lane 等人针对以上问题利用 UNIX 用户的命令数据做了大量实验,以下是他们得出的一些结论.

(1)随着序列长度 l 的增大(从 1～15),在有些情况下检测性能呈下降趋势,有些情况下没有明显变化或轻微上升(和具体用户有关).在聚类、按出现概率提取、按时间顺序截取、随机选择等样本序列提取方法中,聚类方法对不同用户的适应性要强一些,但实现起来最复杂.

(2)关心相邻两命令之间相关性的相似度函数对应的检测性能比不关心相关性的要好一些,最大值按多项式增长和按指数增长的相似度函数对应的检测性能差别不大;总体上看,关心相邻两命令之间相关性且最大值按多项式增长的相似度函数对应的检测性能最好.此外,均值滤噪和中值滤噪算法的性能差别不大.

(3)窗长度越长,检测性能越好;但是窗长度的增大会降低检测的实时性.

3 一种新的基于机器学习的异常检测模型

3.1 变长命令序列检测模型的描述

在用户行为异常检测中,行为模式是指用户操作过程中体现出的某种规律性.实际中,不同用户所具有的行为模式存在差异,同一用户在完成不同行为模式时所执行的命令个数也不尽相同,因而,长度固定的命令序列所组成的样本库一般难以全面准确地反映用户的行为轮廓,这是定长命令序列检测模型的主要缺点.模型的另一缺点在于不容易估算针对具体用户的最佳序列长度;Lane 等人主要采用实验方法来确定最佳序列长度,这种方法所需的计算量很大,而且其性能缺乏稳定性.我们针对定长命令序列检测模型的以上不足进行了改进和修正,提出一种变长命令序列检测模型,具体描述如下:

(1)定义 W 种长度不同的序列,针对每种序列建立一个样本序列库,用 W 个样本序列库来代表一个(或一组)合法用户的正常行为轮廓.按照合法用户命令序列流中各序列的出现概率来提取样本序列.

设序列长度的集合为 $C=\{l(1), l(2), \cdots, l(W)\}$,

其中 $l(i)$表示第 i 种序列的长度,且 $l(1)<l(2)<\cdots<l(W)$. 在样本序列库的个数 W 确定的情况下,C 可有不同的选择. 例如 $W=3$ 时,C 可以为{1,2,3}(即三种序列的长度分别为 1,2,3),也可以为{3,6,9}或其它组合. 设 W 个样本序列库的集合 $L=\{L(1),L(2),\cdots,L(W)\}$,其中 $L(i)$表示长度为 $l(i)$的序列对应的样本序列库. 设一个合法用户的正常训练数据为 $R=(s_1,s_2,\cdots,s_r)$,R 对应的长度为 $l(i)(1\leq i\leq W)$的命令序列流可表示为 $S^i=(Seq_1^i,Seq_2^i,\cdots,Seq_{r-l(i)+1}^i)$,其中 $Seq_j^i=(s_j,s_{j+1},\cdots,s_{j+l(i)-1})$. 我们设定一个概率门限 η,将 $S^i(1\leq i\leq W)$中出现概率大于 η 的命令序列视为合法用户的正常行为模式,$L(i)$即是由这些命令序列组成(一个命令序列的出现概率是指此命令序列在相应命令序列流中的出现次数与该命令序列流中的序列总数之比). 当有多个合法用户时,可以针对每个用户建立 W 个样本序列库,然后根据序列长度将这些用户的样本序列库分别组合在一起构成 W 个总的序列库,用这些序列库来代表这些合法用户整体的正常行为轮廓.

(2)以命令为单位进行相似度赋值,并将加窗滤噪后的相似度输出值作为对被监测用户的行为(或类别)进行判决的依据.

设被监测用户在被监测时间内所执行的 shell 命令流为 $\bar{R}=\{\bar{s}_1,\bar{s}_2,\cdots,\bar{s}_r\}$,它所对应的长度为 $l(i)$的命令序列流为 $\bar{S}^i=\{\overline{Seq}_1^i,\overline{Seq}_2^i,\cdots,\overline{Seq}_{r-l(i)+1}^i\}$,其中 $\overline{Seq}_j^i=(\bar{s}_j,\bar{s}_{j+1},\cdots,\bar{s}_{j+l(i)-1})$. 工作时,按以下方法计算每个 shell 命令 $\bar{s}_j$ 对应的相似度 $\mathrm{Sim}(\bar{s}_j,L)$.

1. 设定 $j:=1,i:=W$.

2. 如果 $j\leq r-l(W)+1$,将 $\overline{Seq}_j^i$ 同 $L(i)$进行比较;否则,结束计算.

3. 如果 $\overline{Seq}_j^i\in L(i)$(即 $\overline{Seq}_j^i$ 与 $L(i)$中的某个序列相同),则 $\mathrm{Sim}(\bar{s}_j,L):=\mathrm{Sim}(\bar{s}_{j+1},L):=\mathrm{Sim}(\bar{s}_{j+2},L):=\cdots:=\mathrm{Sim}(\bar{s}_{j+l(i)-1},L):=2^{l(i)}$,$j:=j+l(i)$,$i:=W$,并返回步 2. 否则,$i:=i-1$.

4. 如果 $i\neq 0$,返回步 2;若 $i=0$,则 $\mathrm{Sim}(\bar{s}_j,L):=0$,$j:=j+1$,$i:=W$,并返回步 2.

按照以上方法进行计算,可得到按时间顺序排列的相似度输出值序列 $Y=(\mathrm{Sim}(\bar{s}_1,L),\mathrm{Sim}(\bar{s}_2,L),\cdots,\mathrm{Sim}(\bar{s}_K,L))$,其中 $r-l(W)+1\leq K\leq r$,对这个相似度输出值序列进行加窗滤噪处理,得如下相似度输出值:

$$N(i,L)=\frac{1}{w}\sum_{j=i-w+1}^{i}\mathrm{Sim}(\bar{s}_j,L)\qquad(3)$$

式中 $N(i,L)$表示被监测用户命令流 $\bar{R}$ 中第 i 个命令时刻对应的相似度输出值,w 表示窗长度;i 的初值为 w,其增长步长为 1. w 是一个重要参数,它决定了从被监测用户行为发生到检测系统对其行为(或类别)做出判断的最短时间(即检测时间). 在不考虑计算时间的情况下,检测时间为 w 个命令持续时间.

对 $N(i,L)$设定一个门限 λ,若它大于 λ,将被监测用户的当前行为判为正常行为(或将被监测用户判为合法用户),否则,将其判为异常行为(或将该用户判为非法用户).

(3)当有了新的合法用户正常训练数据时,可以根据新数据重新计算各序列的出现概率,进而对样本序列库作出调整. 因此,该模型对合法用户正常行为的变化是具有适应性的.

3.2　模型的分析与比较

同 Lane 等人的定长序列检测模型相比,我们提出的变长序列检测模型有以下特点:

(1)认为用户的不同行为模式所对应的命令序列长度是不同的,建立多个序列长度不同的样本序列库来描述合法用户的正常行为轮廓. 这一假设更符合一般用户的实际情况,因而所建的样本序列库能够更好地反映合法用户的正常行为轮廓.

(2)以指令为单位进行相似度赋值. 赋值方法为:以当前命令为起点组成多个长度不同的序列,并按照长度从大到小的顺序依次同相应的样本序列库进行比较,如果其中一个序列同相应样本序列库中的某个序列相同,则将此序列中每个命令所对应的相似度赋以相同的值,序列长度越长,所赋的值也越大(如果任何一个序列同相应样本序列库中的序列都不相同,则将当前命令对应的相似度赋以零值);然后再以此序列(或命令)之后的下一个命令为起点组成多个序列重新进行上述赋值过程. 这种赋值方法的实质是在被监测用户命令流中进行行为模式挖掘,它主要关心被监测用户当前命令序列所代表的行为模式是否能够同合法用户的某个正常行为模式完全匹配,不存在 Lane 等人赋值方法中的模糊性.

(3)由于需要建立多个样本序列库,此模型在描述合法用户正常行为时对检测系统存储空间的需求较大. 但是,以指令为单位赋值的方法使得模型工作时需要的计算量要小一些,这可以提高检测的实时性.

4 实验结果

我们对上述两种模型的性能进行了实验,实验数据为普渡大学公开发布的 8 个用户两年时间内在 UNIX 平台上执行的 shell 命令,实验中我们采用了其中的 4 个用户 USER1,USER2,USER3,USER4 的数据,其中每个用户各有 15000 个命令.每个用户的前 10000 个命令用于模型的训练(建立样本序列库),后 5000 个指令用于模型的性能测试.实验共分 4 组,每组实验选取一个用户作为合法用户(将其行为设为正常行为),而将其他 3 个用户设为非法用户(将其行为设为异常行为).在定长序列检测模型中,序列长度 $l=3$,窗长度 $w=91$;在变长序列检测模型中,序列长度集合 $C=\{1,2,3\}$,窗长度 $w=91$.两种模型中,各个样本序列库均由合法用户近 10000 个序列中出现概率大于 0.02%的序列组成.

图 2 给出了 USER1 设为合法用户的情况下,采用定长序列检测模型时 USER1 和 USER2 的归一化相似度输出曲线,图 3 给出了采用变长序列检测模型时相应的归一化相似度输出曲线(为绘图方便,对横坐标做了平移).可以看出,图 3 中 USER1 对应的相似度输出曲线同 USER2(非法用户)对应的相似度输出曲线的可分性明显好于图 2,而且,图 3 中 USER1 的相似度输出曲线的局部抖动也相对小一些.

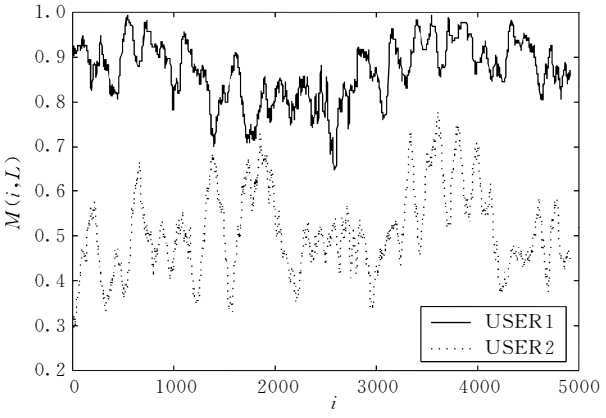

图 2 定长序列模型的相似度输出曲线

图 4 给出了由实验结果得到的 ROC 曲线,图中横坐标是 4 组实验中对合法用户正常行为的平均虚警概率,纵坐标为对非法用户异常行为的平均检测概率,model-1 表示定长序列检测模型,model-2 表示变长序列检测模型.由图可见,变长序列检测模型的性能明显好于定长序列检测模型的性能.

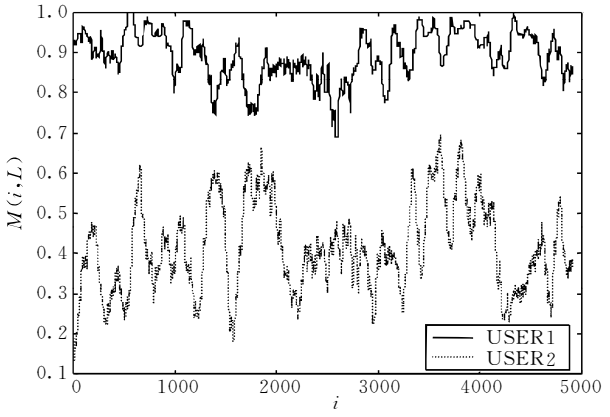

图 3 变长序列模型的相似度输出曲线

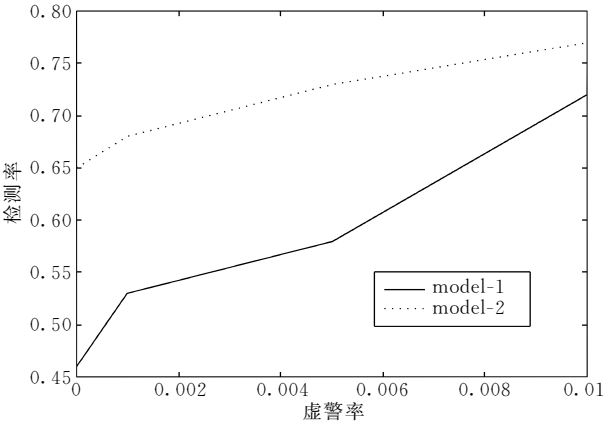

图 4 两种模型的 ROC 曲线

5 结束语

本文提出了一种基于机器学习的 IDS 变长命令序列检测模型,并同 Lane 等人的定长序列检测模型进行了对比.基于 UNIX 用户命令数据的实验结果表明,同等虚警概率条件下此模型有更高的检测概率.根据 Lane 等人的分析,定长序列检测模型的学习方法和检测算法对不同的审计数据有一定的适应性,因而变长序列检测模型也应该可以用于以系统调用为审计数据的程序行为异常检测,但具体的检测性能还有待进一步地研究和实验.

参 考 文 献

1 Lane T,Brodley C E. An application of machine learning to anomaly detection. In:Proceedings of the 20th National Informa-

tion Systems Security Conference, Baltimore Marylard, USA, 1997. 366～377

2 Kosoresow A P, Hofmeyr S A. A shape of self for UNIX processes. IEEE Software, 1997, 14(5): 35～42

3 Lee W, Stolfo S J. Data mining approaches for intrusion detection. In: Proceedings of the 7th USENIX Security Symposium, San Antonio, Texas, USA, 1998. 66～72

4 ISS. Network-Vs. Host-Based Intrusion Detection, 1998

5 Lane T. Machine learning techniques for the computer security domain of anomaly detection[Ph D dissertation]. Purdue University, West Lafayette, IN, 2000

6 Zhao Hai-Bo, Li Jian-Hua, Yang Yu-Hang. Network intrusion intelligent real-time detection system. Journal of Shanghai Jiaotong University, 1999, 133(1): 76～79(in Chinese)
(赵海波,李建华,杨宇航. 网络入侵智能化实时检测系统. 上海交通大学学报, 1999, 133(1): 76～79)

SUN Hong-Wei, born in 1964, Ph. D. candidate. His research interests include information security, intrusion detection.

TIAN Xin-Guang, born in 1976, Ph. D. candidate. His research interests include digital signal process, information security.

LI Xue-Chun, born in 1963, professor. His research interests include information security, mobile communication.

ZHANG Er-Yang, born in 1941, professor. His research interests include digital signal process, communication network.