南京航空航天大学

论文集

（二〇〇七年） 第19册

信息科学与技术学院

（第3分册）

南京航空航天大学科技部编

二〇〇八年三月

信息科学与技术学院

044～045 系

信息科学与技术学院2007年发表论文目录-4.1

序号	姓名	职称	单位	论文题目	刊物、会议名称	年、卷、期	类别
1	张有东 王建东 庄　毅	博士生 教　授 教　授	044 044 044	基于m序列的隐私保护数据擦除系统设计	小型微型计算机系统	2007.28.05.	
2	张有东 王建东 朱梧槚	博士生 教　授 教　授	044 044 044	反计算机取证技术研究	河海大学学报(自然科学版)	2007.35.01	
3	张有东 任勇军 王建东	博士生 博士生 教　授	044 044 044	Network Forensic Computing Based on ANN-PCA	2007年IEEE International Conference on Computational Intelligence and Security会议交流		
4	任勇军 王建东 庄　毅 方黎明	博士生 教　授 教　授 博士生	044 044 044 044	动态的Ad Hoc网络密钥管理方案	南京航空航天大学学报	2007.39.06	
5	任勇军 王建东 庄　毅	博士生 教　授 教　授	044 044 044	基于双线性对的Ad hoc网络主动秘密共享方案	系统工程与电子技术	2007.29.09	
6	任勇军 王建东 庄　毅 方黎明	博士生 教　授 教　授 博士生	044 044 044 044	基于身份的Ad Hoc网络门限密钥发送协议	武汉大学学报(工学版)	2007.40.05	
7	任勇军 王建东 张有东 方黎明	博士生 教　授 博士生 博士生	044 044 044 044	Identity-Based Key Issuing Protocol for Ad Hoc Networks	2007年IEEE International Conference on Computational Intelligence and Security会议交流		
8	李　涛 王建东 叶飞跃 冯新宇 张有东	博士生 教　授 博士生 博士生 博士生	044 044 044 044 044	一种基于用户聚类的协同过滤推荐算法	系统工程与电子技术	2007.29.07	
9	李　涛 王建东 叶飞跃	博士生 教　授 博士生	044 044 044	推荐系统中一种新的相似性计算方法	计算机科学	2007.34.08	
10	朱俊武 王建东 李　斌	博士生 教　授	044 044 外	SP4IAIS:A Semantic Platform for Integrating Aviation Information systems	2007年International Conference on Convergence Information Technology (2nd ICCIT'07)会议交流		
11	朱俊武 王建东 李　斌	博士生 教　授	044 044 外	A Cooperation Infrastructure Based on Semantic and Ist Application to Civil Aviation Management	2007年International Conference on Convergence Information Technology (2nd ICCIT'07)会议交流		
12	朱俊武 王建东 李　斌	博士生 教　授	044 044 外	SSOA;a Semantic Service-Oriented Architecture Based on Fuzzy Assertion System	2007年International Conference on Cooperative Work in Design(11th CSCWD'07)会议交流		
13	朱俊武 王建东 李　斌	博士生 教　授	044 044 外	On Dynamic and Concurrent Model of Web Service Components	2007年International Conference on Supported Cooperative Work in Design 11th CSCWD'07)会议交流		

信息科学与技术学院2007年发表论文目录-4.2

序号	姓名	职称	单位	论文题目	刊物、会议名称	年、卷、期	类别
14	庄　毅 刘　坤 陈　尉 陆佳琦	教　授 硕士生 硕士生 硕士生	044 044 044 044	一种基于Agent的主动式信息安全模型	南京理工大学学报(自然科学版)	2007.31.01	
15	朱丽丽 庄　毅 叶延凤 甘春闰	硕士生 教　授 副教授 硕士生	044 044 044 044	虚拟装配中快速碰撞检测算法的研究与实现	计算机应用	2007.27.12	
16	邹玉娟 庄　毅 张　霞 朱　勐	硕士生 教　授 硕士生 硕士生	044 044 044 044	某型投物伞系统三维动态仿真的设计与实现	计算机仿真	2007.24.06	
17	夏正友	副教授	044	Using the loopy belief propagation in siguo	ICGA Journal	2007.30.04	
18	夏正友	副教授	044	Emergence of Social Rumor:Modeling,	Lecture Notes in	2007.4490.00	
19	陆　慧 夏正友	硕士生 副教授	044 044	四国军棋游戏中搜索算法的实验与分析	江南大学学报	2007.12.00	
20	陆　慧 夏正友	硕士生 副教授	044 044	基于两种策略的四国军棋评价函数的实验与分析	全国人工智能大会	2007.00.00	
21	林　菡 夏正友	硕士生 副教授	044 044	TAC-SCM计算机博弈系统设计与分析	江南大学学报	2007.12.00	
22	袁家斌 顾恺恺 姚　黎	副教授 硕士生 硕士生	044 044 044	基于信息安全控制原理的安全网络技术	南京理工大学学报(自然科学版)	2007.31.04	
23	姚　黎 袁家斌	硕士生 副教授	044 044	基于信息安全控制的安全网络设计	福建电脑	2007.00.02	
24	叶　峰 袁家斌	硕士生 副教授	044 044	置换密钥矩阵加密算法安全性分析	南京航空航天大学学报	2007.39.06	
25	何安元 袁家斌	硕士生 副教授	044 044	基于组合学的大规模磁盘队列轮换休息机制的编码设计	福建电脑	2007.05.00	
26	何安元 袁家斌	硕士生 副教授	044 044	输媒体流—设计和部署3TNet南京航空航天大学驻地网	中国教育网络	2007.00.03	
27	陈文彬 孟江涛	讲　师 讲　师	044 043	Approximation Algorithms for k-Duplicates Combinatorial Auctions with Subadditive Bidders	Lecture Notes in Computer Science	2007.4616	
28	陈文彬 孟江涛	讲　师 讲　师	044 043	An Improved Lower Bound for Approximating Shortest Integer Rrlation in infinity norm	Information Processing Letters	2007.101	
29	陈文彬 孟江涛	讲　师 讲　师	044 043	Hardness of Approximating the Minimum Solution of Linear Diophantine Equation	Theoretical Computer Science	2007.374	

信息科学与技术学院2007年发表论文目录-5.1

序号	姓名	职称	单位	论文题目	刊物、会议名称	年、卷、期	类别
1	朱小栋 黄志球 王　磊 程　亮 沈国华	博士生 教　授 硕士生 硕士生 讲　师	045 045 045 045 045	基于粒计算的时序数据流关联规则挖掘模型研究	计算机科学	2007.34.8A	
2	朱小栋 黄志球 张君华 杨淑群	博士生 教　授 博士生 博士生	045 045 045 045	Granular Computing based Intrusion Detection Model upon Network Monitor Data Streams	2007年2nd International Conference on Pervasive Computing and Applications会议交流		
3	朱小栋 黄志球 杨淑群 沈国华	博士生 教　授 博士生 讲　师	045 045 045 045	Fuzzy implication methods in fuzzy logic	2007年Fourth International Conference On Fuzzy Systems And Knowledge Discovery会议交流		
4	赵晓非 黄志球	博士生 教　授	045 045	基于CWM的元数据的形式化推理框架研究	计算机研究与发展	2007.44.05	
5	凌兴宏 黄志球	博士后 教　授	045 045	结合逻辑和决策论方法的Agent模型研究	南京航空航天大学学报	2007.39.06	
6	祝　义 黄志球	博士生 教　授	045 045	UML与Z结合的建模过程及其应用	计算机科学	2007.34.05	
7	袁　敏 黄志球	博士生 教　授	045 045	一种主动式有状态的Web时效索引机制	计算机应用研究	2007.24.07	
8	张联超 黄志球 沈国华 周　航	硕士生 教　授 讲　师 博士生	045 045 045 045	基于逆向工程的本体构建方法研究	计算机工程与设计	2007.28.24	
9	程　亮 张联超 黄志球	硕士生 硕士生 教　授	045 045 045	基于语义的XML数据清洗框架	郑州大学学报	2007.39.04	
10	沈国华 黄志球 朱小栋	讲　师 教　授 博士生	045 045 045	Using Description Logics Reasoner forOntology Matching	2007年the Workshop on Intelligent Information Technology Application (11TA'07)会议交流		
11	沈国华 张育平 黄志球	讲　师 副教授 教　授	045 043 045	在工程实践中培训软件能力	南京航空航天大学学报（社会科学版）	2007.09.01	
12	皮德常 秦小麟	副教授 教　授	045 043	Technologies about Mining Tendency Association Rule	Journal of Information and Computational Science	2007.04.00	
13	崔延良 皮德常	硕士生 副教授	045 045	SQLmmbd:An Embedded Main Memory Database Management System	Information Technology Journal	2007.06.06	
14	袁培森 皮德常	硕士生 副教授	045 045	用于内存数据库的Hash索引的设计与实现	计算机工程	2007.33.18	

小 型 微 型 计 算 机 系 统
Journal of Chinese Computer Systems

2007 年 5 月 第 5 期
Vol. 28 No. 5 2007

基于 m 序列的隐私保护数据擦除系统设计

张有东[1,2],王建东[1],庄 毅[1]

[1](南京航空航天大学 计算机科学与工程系,江苏 南京 210016)
[2](淮阴工学院 计算机工程系,江苏 淮安 223001)
E-mail:z.yd@163.com

摘 要:网络技术的发展使得隐私保护日益引起关注.本文分析比较了数据摧毁和数据擦除技术在隐私保护方面的应用,阐述了数据擦除技术的发展现状.结合现代 RLL 码的发展,提出了一种基于 m 序列的数据擦除算法,并介绍了一个数据擦除系统设计的主要技术.

关 键 词:数据摧毁;数据擦除;RLL 码;m 序列
中图分类号:TP309 **文献标识码**:A **文 章 编 号**:1000-1220(2007)05-0826-04

Data Wiping System Design for Privacy Protecting Based on m Sequence

ZHANG You-dong[1,2], WANG Jian-dong[1], ZHUANG Yi[1]

[1](*Nanjing University of Aeronautics and Astronautics, Department of Computer Science and Engineering, Nanjing* 210016, *China*)
[2](*Huaiyin Institute of Technology, Department of Computer Engineering, Huaian* 223001, *China*)

Abstract: With the development of networking technology, privacy protection is more and more important. The paper analysis and compares the application of the data destroy and data wiping technology in the privacy protection domain, and illustrates the actuality of the data wiping technology. Since the development of the modern RLL code technology, the paper proposals a data wiping algorithm based on m sequence, and introduces the main technology to design a data wiping system.

Key words: data destroy; data wiping; RLL code; m sequence

1 引 言

现代计算机操作系统和应用程序在运行时会产生大量关于用户活动的记录,这些记录越来越成为法律调查和用户个人关注的焦点,特别是在共享环境下存在的对个人隐私安全的风险.具统计,仅美国每年由于网络隐私引起的经济损失已达数百亿美元.因此,对个人隐私的保护已成为一个急待解决的有关社会、法律和技术的综合性问题[1],个人隐私的保护问题已引起广泛关注,相关方面的研究也越来越引起重视.数据擦除(Data Eliminating or Data Wiping)技术则是目前隐私保护中常用的一种技术.

2 隐私保护与数据摧毁

隐私是一种个体决定和控制自己的信息被他人共享或使用的权利.随着计算机特别是网络技术的发展,在个人计算机中保存了大量的电子数据,在 ISP 的存储中保留了大量用户提交的信息,另外,很多软件从方便用户的角度也暂存了用户最近的活动记录,这些都使得用户控制自己的信息被他人共享或使用的能力受到极大挑战,个体隐私受到严重威胁.

保护隐私最直接的方法是删除文件或数据,清除软件使用后的痕迹,如上网后清除 Internet 临时文件夹中的所有文件,但是,这种删除操作对于数据恢复工具来说很容易恢复.通过对丢弃的硬盘的分析,研究人员在恢复的数据中发现了银行交易记录、公司数据等大量隐私信息,1986 年发生的 Envon Corp 调查案使人们开始意识到,删除文件并不意味着文件所包含的信息的湮没(Obliterate)[2],即使是对删除的数据进行了覆盖或对磁盘进行了格式化.

事实上,诸如磁力显微镜(Magnetic Force Microscopy, MFM)和扫描隧道显微镜(Scanning Tunneling Microscopy, STM)技术可以从磁介质中找到被完全重写的数据;内存芯片也有一种未公开的诊断模式,允许访问比 1 比特位小的信息,通过修改电路,类似于信号调制,可以检索旧的信息[3].面对这些技术,真正地从磁介质上删除数据是非常困难的.

保护个体隐私的有效方法之一是摧毁电子数据,Peter Gutmann 在 1996 年就提出了数据摧毁的问题[4],从存储原理上讲,数据摧毁可以有三种选择:摧毁实际数据,摧毁元数据(meta-data),或者是两者的结合.对于磁存储介质上的数据摧毁,有物理摧毁和逻辑摧毁两种方法,物理摧毁包括消磁(degaussing)或暴力(brute force)破坏,逻辑摧毁主要是数据擦除方法,目前的一些工具软件提供了数据擦除功能,如 Norton System Works 中的 WipeInfo 功能,也有一些专用的软件,如 Eraser, FileWipe 等,其实现主要是基于简单的 0、1

收稿日期:2006-03-16 基金项目:国家重点基础研究发展规划"九七三"基金项目(G1999032701)资助. 作者简介:张有东,男,1967 年生,博士研究生,副教授,研究方向为数据挖掘、入侵检测;王建东,男,1945 年生,教授,博士生导师,主要研究方向为人工智能、知识工程、机器学习、数据挖掘等.

交替覆盖方案,测试表明,大多数磁盘擦除工具并不能完全清除所有数据,仅比简单的删除多了一点保护[2].

3 数据擦除技术分析

数据擦除技术用于定位用户的活动记录,搜寻并不可逆地删除它们,即所谓的数据安全删除,目的是使数据恢复无法进行.数据擦除的基本思想是针对磁介质存储的特性,用某种方案反复重写磁介质,DOD 5220.22-M(C and E)是美国国防部提出的标准擦除算法,该标准推荐 3 次 0、1 交替覆盖方案,美国国家安全局(NSA)则推荐进行 7 次覆盖方案[5].事实上,数据擦除方案的设计与磁表面编码方式有关,由于写入电流的幅度、相位、频率变化不同,编码方式的形式也有多种,主要包括 RZ、NRZ 和 NRZ1 方式,FD、PE、MFM、MDM(M2FM)按位编码方式,GCR(4/5)和 PM 成组编码方式.

编码方式的改进是提高磁盘存储密度的重要技术之一,现代高密度磁盘主要采用 RLL 码(Run Length Limited Code),RLL 码是编码理论中研究码制变换、增强抗干扰能力而得出的一种编码,其实质是将编码前的 m 位原始数据序列变为零的个数受限的 n 位记录序列(长度为 n 的游程),然后再用 NRZ1 制方式进行调制和写入.

Gutmann 针对当时的三种磁盘编码 MFM、RLL(2,7)(PWM)和 RLL(1,7)(PPM)提出了一种擦除方案,共需要进行 35 次覆盖,它被认为是最安全的擦除算法[4,6].其覆盖方案如表 1 所示.

表 1　Gutmann 覆盖方案

Table 1　The Gutmann's overwriting scheme

次数	重写的数据	编码方案
1～4	Random	
5	01010101 01010101 01010101 0x55	(1,7)RLL　　MFM
…	…	… … …
31	11011011 01101101 10110110 0xDB 0x6D 0xB6	(2,7)RLL
32～35	Random	

Gutmann 方案针对三种编码数据恢复的电磁学原理,采用一组周期性的数据反复重写磁介质,以彻底擦除磁盘存储器中的数据.

实际上,可用五元组(d,k;m,n;r)来表示不同的 RLL 码,其中,d、k 分别表示记录序列中相邻两个 1 之间至少有 d 个零、最多不得超过 k 个零,r 为一次变换时,最大和最小长度之比,这样,前述编码方式都可以用 RLL 码来表示,如:PE(0,1;1,2;1)、MFM(1,3;1,2;1)、GCR(4/5)(0,2;4,5;1)、3PM(2,11;3,6;1)等.

Gutmann 方案的缺点是速度太慢,只对 MFM、PWM 和 PPM 具有较好的安全性.而现代编码表示方法的统一,使得正确设计 d,k 值,即可获得优良的 RLL 编码性能,因而在近几年发展的高密度磁盘中不同的 RLL 码得到了广泛应用,如希捷 Momentus 硬盘采用了 RLL(0,11)游程编码等.为此,我们设计了用 m 序列来模拟产生通用覆盖序列的覆盖方案 GRLLs(General RLL Scheme),并针对 Windows 的 FAT32 文件结构用 VC++6.0 实现了该方案,测试表明,该方案既提高了擦除速度,又保证了数据的不可恢复性.

4 隐私保护数据擦除系统设计主要技术

4.1 文件搜索定位

文件搜索定位主要是搜索待擦除文件的 FAT 信息,定位该文件在磁盘数据区的起始簇号,形成数据存储的簇链.下面针对 FAT32 仅给出程序设计的主要描述和解释:

S1:读出被检索文件的 FRI;

S2:InitClouster = = * (FRI+1aH);

S3:FirstSectorAddress = = (InitClouster-2) * 16;

S4:ClousterOffset = = InitClouster * 4 mod 512;

S5:SecondClouster = = * (FAT+ClousterOffset);

S6:if (SecondClouster = = 0XFFFFFF0FH) exit;

S7:else InitClouster = = SecondClouster;

S8:goto S3;

FRI 为目录登记项,一个目录登记项占用 32B,其字节位移 1aH 开始的一个字的值记录的是文件的起始簇号.FirstSectorAddress 是文件在 DATA 区内第一簇首扇区的存储地址,这个地址是从逻辑驱动器的 DATA 区首扇区开始计算的,如果从硬盘的起始扇区开始计算,还要加上前面所有的扇区数,16 为 FAT32 的每簇扇区数,可从 BPB 表中得到;ClousterOffset 为该簇登记项在 FAT 表中的字节位移;FFFFFF0FH 是 FAT32 簇结束标志.

4.2 m 序列生成算法

m 序列是一种二进制伪随机序列,通常由线型反馈移位寄存器生成,设一线型反馈移位寄存器为 r 级,其序列多项式 G(x)可表示为:

$$G(x)=a_0+a_1x+a_2x^2+\cdots+a_nx^n+\cdots=\sum_{n=0}^{\infty}a_nx^n$$

其中,a_n 是一个二元有限域{0,1}的元素,运算为模 2 和,显然,序列$\{a_n\}$与 G(x)一一对应.再根据 Caley-Hamiton 定理,可推出其特征多项式 f(x)为:

$$f(x)=\sum_{i=0}^{r}c_ix^i$$

其中，c_i 为反馈系数，其取值也是一个二元有限域{0,1}的元素.通过变换，可推出序列多项式与特征多项式之间的关系为：

$$G(x)=x/f(x)$$

由此可得到产生序列$\{a_n\}$的方法，即对 f(x)进行长除，得到序列多项式 G(x)，G(x)的系数就是所求序列.当该线性反馈移位寄存器的抽头系数取自一本原多项式时，所得序列即为一 m 序列.

根据 m 序列的性质，其周期 $L=2^r-1$，考虑到 1 个扇区的字节数为 512，若 L+1 对应于一个磁盘扇区字节数，则可得 r 为 9，其本原多项式共有 $\Phi(2^9-1)/9=48$ 条（Φ 为 Euler 函数），可从中选择反馈系数(1000010001)，即得本原三项式：

$$f(x)=x^9+x^4+1$$

其产生的 m 序列长度为 511 位，在最后加上与其前一位反转的值，就得到一个 512 位的序列 m_1.

m_1 可以用软件方法产生，编程时，可根据 RLL(d,k)码的性质，在产生 m_1 时对两个 1 之间零的个数进行限制，即两个 1 之间至少有 d 个、最多不得超过 k 个零.这只需要在算法中增加 0 计数器 Count0，并作如下判断：当产生 1 个 1 时 Count0 开始对 0 计数，当下一个 1 产生时，若 Count0<d，则将该 1 反转，同时将 Count0 加 1；当下一个 1 产生前，若 Count0>k，则将该 0 反转，并重置 Count0.到此为止，就产生了一个针对不同 RLL 码的以扇区为单位的磁翻转序列 m_2，而且，不同于流加密，这样产生的序列是不可逆的.

综上所述，我们可以根据不同的 RLL 码，产生一个以扇区为单位的、具有密码学性质且不可逆的覆盖序列 m_2，一组 $m_2^i(i=1\cdots n)$即可组成一个覆盖方案，其覆盖次数 n 可根据用户对删除数据的重要性来确定.当然，m 序列也可用 M 序列来代替，此时可直接产生周期为 2^r 的序列，而无须在最后补一位.

设 a[m_SIZE]为 r 级寄存器数组，m2_sequence[m_SIZE+1]为产生的 m_2 序列，⊕表示模 2 和运算，且，m_SIZE=510，r=9，本系统中 m 序列的实现算法描述如下：

```
Count0==-1;
for (i=0;i<m_SIZE;i++)
  {m2_sequence[i]==a[0];
   if (m2_sequence[i]==1) Count0++;
   for (j=0;j<m_SIZE;j++)
     a[j]==a[j+1];
   a[m_SIZE]==a[9]⊕a[4]⊕a[0];
   if (m2_sequence[i]==1 & Count0<d)
       {m2_sequence[i]==0;Count0 ++;}
   if (m2_sequence[i]==0 & Count0>k)
       {m2_sequence[i]==1;Count0==-1;}
  }
m2_sequence[m_SIZE+1]==! m2_sequence[m_SIZE];
```

4.3 数据写入覆盖

Windows2000 已不再支持 INT13H 调用，对磁盘等硬件设备的访问通过 CreateFile()函数提供的文件访问机制实现.CreateFile()打开的是整个磁盘逻辑分区，需要通过 SetFilePointer()函数以文件操作的方式把指针移到要操作的磁盘扇区开始处，该函数的参数 hFile 为 CreateFile()返回设备文件句柄，参数 lDistanceToMove 和 lpDistanceToMoveHigh 指出所设置偏移量的低端和高端部分，dwMoveMethod 指文件指针从何处开始移动.在定位到要访问的扇区开始位置后就可以通过 WriteFile()函数实施相应的写入覆盖操作.据此可设计出写处理函数 OverWriteSectors(BYTE bDrive, DWORD dwStartSector, WORD wSectors, LPBYTE lpSectBuff)，实现对磁盘扇区数据的写入.下面是这部分功能的主要实现代码：

```
//写入函数
BOOL CDirectAccessHDDlg::WriteSectors(BYTE bDrive, DWOR
dwStartSector, WORD wSectors, LPBYTE lpSectBuff)
{
HANDLE hDev = CreateFile(devName, GENERIC_WRITE
FILE_SHARE_WRITE, NULL, OPEN_EXISTING, 0, NULL)
if (hDev == INVALID_HANDLE_VALUE) return 0;
SetFilePointer(hDev, 512 * dwStartSector, 0, FILE_BEGIN);
DWORD dwCB;
BOOL bRet = WriteFile(hDev, lpSectBuff, 512 * wSectors
&dwCB, NULL);
CloseHandle(hDev);
return bRet;
}
...
//按扇区用 m2 序列覆盖
while (search_flags){
memset(bBuf, m2_sequence, sizeof(bBuf));
bRet = OverWriteSectors(uDiskID, i, 1, bBuf);
}
```

5 性能分析与实验

由于数据擦除的程度与覆盖次数及磁盘记录编码有关，这是由磁盘存储的磁场原理决定的，从前面的分析可见，本文所提出的方案具有三个性质：

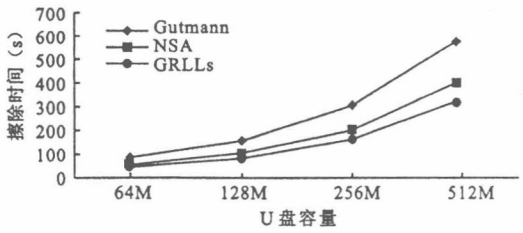

图 1 三种方案擦除速度对比

Fig. 1 Erasing speed comparing of three schemes

(1) 是一种基于 RLL 码原理的通用覆盖方案；

(2) 覆盖是按扇区进行的；

(3) 覆盖序列具有密码学性质.

因此，本文所提出的方案，对于相同的覆盖次数而言，

据性质(2)(3),其安全性能和速度高于NSA方案,如果覆盖次数相同,根据性质(1)(2),其速度高于Gutmann方案且具有通用性.

下面是用VC++6.0实现三种方案,在WindowsXP平台下的比较实验结果,擦除对象为LX 64M、128M、256M、512M四款U盘(USB2.0接口)上的全部内容,覆盖次数为三次.

由于磁盘的读写速度、文件FAT结构、文件搜索定位时间、文件块大小、磁盘碎片等因素都会对文件擦除的速度产生影响,为了便于比较分析,本实验采用对U盘的完全擦除方法以避免上述因素的影响.另外,测试中对Gutmann方案采用了通用的seed算法产生1~3次的重写数据.

6 结束语

本文分析比较了数据摧毁和数据擦除技术在隐私保护方面的应用,系统阐述了数据擦除技术的现状与发展,并结合现代RLL码的发展,提出了一种用m序列来模拟产生通用覆盖序列的一种数据擦除算法.同时也应注意到,虽然随着人们对隐私保护需求的日益扩大,数据擦除技术的应用也日益广泛,但同其它网络安全技术一样,技术的发展总有其多面性,这种新的技术也可能被用于反计算机取证,这是在进一步研究数据擦除技术时不能不考虑的因素.

References:

[1] Abdelmounaam Rezgui, Athman Bouguettaya, Mohamed Y Eltoweissy. Privacy on the Web: facts, challenges, and solutions [J]. IEEE Security & Privacy, 2003,(6):40-49.

[2] Matthew Geiger. Evaluating commercial counter-forensic tools [C]. DWFS 2005. New Orleans, LA, August, 2005.

[3] Farmer Venema. The persistence of deleted file information [M]. PWS Publishing Company, 2004.

[4] Peter Gutmann. Secure deletion of data from magnetic and solid-state memory[C]. In: the Sixth USENIX Security Symposium Proceedings, San Jose, California, July, 1996.

[5] U.S. Department of Defense. Standard 5220.22-M: national industrial security program operating manual (NISPOM), Chapter8[S/OL]. http://www.dss.mil/isec, January, 1995.

[6] Peter Bennison. Data security for surplus or end of life hardware [C]. 2004 NYS Cyber Security Conference: New york, April 21-22, 2004.

反计算机取证技术研究

张有东[1,2],王建东[1],朱梧槚[1]

(1.南京航空航天大学计算机科学与工程系,江苏 南京 210016; 2.淮阴工学院计算机工程系,江苏 淮安 223003)

摘要:分析了反计算机取证的基本概念和方法,比较了反计算机取证所采用的数据擦除、数据加密、数据隐藏、数据混淆和数据转换等主要技术,并提出了一种安全、高效的基于 m 序列的数据擦除方法.

关键词:反取证;数据擦除;数据隐藏;数据混淆;数据转换

中图分类号:TP311 **文献标识码**:A **文章编号**:1000-1980(2007)01-0104-04

目前,在计算机网络犯罪手段与网络安全防御技术之间的对抗不断升级的形势下,计算机取证作为一门交叉学科,越来越受到计算机安全和法律专家的重视.与此同时,一种新的反计算机取证(Anti-Computer Forensic,以下简称反取证)技术也正悄然兴起.国内外关于这种技术公开的、系统的报道还不多.然而,通过分析研究反取证技术以保证计算机取证的科学性和有效性是非常重要的[1].本文将提出反取证技术的定义与方法,分析比较反取证采用的主要技术,并提出一种基于 m 序列的反取证算法.

1 反取证定义

黑客技术、反侦查技术和入侵检测逃避的主要目的是逃避系统的安全检测以成功地侵入系统,但同时也会在系统中留下大量的电子证据.随着取证技术的发展,高明的计算机犯罪分子在作案前开始周密计划和做反跟踪准备,作案后更改、删除目标机的信息,并清理自己的工具机,这是反取证技术兴起的最初动因.另外,反取证技术的发展还来源于保护个人隐私的需要,即保护合法公民或公司的正当法律权利不受侵犯.

早期的反取证技术可定义为"删除或破坏电子证据,使得取证方法无效".随着计算机取证的程序化及其技术的完善,反取证发展为针对计算机取证过程的各个阶段及其形成证据链(Chain of Custody)的条件,破坏电子证据的调查、保护、收集、分析和法庭诉讼,减少被获取证据数量,降低被获取证据质量,从而尽量隐藏或不在对方系统甚至自己系统中留下法律意义上的证据.

2 反取证基本方法分析

传统的反取证方法有改变文件扩展名、使用 Swap 空间等,现代计算机取证工具已经能很好地处理它们.由于计算机取证调查受技术和非技术因素的影响,反取证的基本方法同样有策略上的考虑和技术上的方法2个方面.

计算机取证需要建立有效的证据链,否则将产生法律疑点(Legal Doubt),也就是使得控方律师在被告有罪的推理中出现疑点,从而不被法庭采纳.由于 Internet 的特点,匿名访问、匿名账号、匿名存储以及利用 Root Kit 绕开系统日志等,都将使取证调查人员在确定罪犯在特定计算机上的活动时产生困难,罪犯能够匿名的越多,调查中的疑点就越多,甚至罪犯会"黑客"自己的计算机,以造成他们的计算机是被其他罪犯利用来犯罪的假象.

在技术层面上,目前的反取证技术主要是针对证据的收集和分析.罪犯利用取证调查开始时影响判断的因素来干扰、阻挠取证,导致调查偏离犯罪活动场所,这些因素包括数据集的信息异常、失效或不产生指定的结果.证据收集阶段的反取证技术主要有摧毁、隐藏、假造信息以及防止证据的创建等.

收稿日期:2006-03-12

基金项目:国家重点基础研究发展规划(973)项目(G1999032701)

作者简介:张有东(1967—),男,江苏淮安人,副教授,博士研究生,主要从事数据挖掘、入侵检测研究.

在取证分析阶段，取证人员已获得并保护了证据，并且将要去理解这些证据，针对这个阶段的反取证技术是通过数据加密阻止分析人员对所获取的证据的理解，甚至通过某种方法歪曲取证人员可能会感兴趣的数据[2]。

另外，由于目前的计算机取证主要由一些取证工具软件完成，反取证的一种实用的方法是针对取证工具软件的缺点进行攻击，如对 Unix 下的取证工具 TCT 的攻击等。

3　反取证主要技术分析

3.1　数据摧毁技术

摧毁证据是阻止取证的最有效的方法，主要攻击取证的收集阶段。从隐私与敏感数据保护角度，Peter Gutmann 在 1996 年就提出了数据摧毁的问题[2]。从存储原理上讲，摧毁证据可以有 3 种选择：摧毁实际数据，摧毁元数据（Meta-Data），或者是两者的结合。对于磁存储介质上的数据摧毁，有物理摧毁和逻辑摧毁 2 种方法，物理摧毁包括消磁（Degaussing）或暴力（Brute Force）破坏，逻辑摧毁则包括改变驻留媒体的数据组成、移走相关数据的踪迹（Traces）等。进一步地，反取证不仅摧毁实际的数据，还可能根除摧毁痕迹，这极大地降低了取证人员在计算机系统上发现证据的能力。

值得注意的是，一般的文件删除甚至删除后的多次覆盖，并不意味着数据的摧毁，事实上，诸如 MFM（Magnetic Force Microscopy）和 STM（Scanning Tunneling Microscopy）技术可以从磁介质中找到被完全重写的数据；内存芯片也有一种未公开的诊断模式，允许访问比 1 比特位小的信息，通过修改电路，类似于信号调制，可以检索旧的信息[3]。

3.2　数据加密技术

虽然摧毁所有潜在的证据很有效，但罪犯可能想保存一些有用的数据。为了使这些数据不被取证人员所理解，罪犯通常会使用加密技术。现代加密技术的发展使得这种方法很容易被犯罪分子所利用，当然也有专门研究用于反取证的加密技术，文献[4]就介绍了一种基于两层 PBKDF2 的反取证数据存储方法。

3.3　数据隐藏技术

数据隐藏（Data Hiding）是一种有效的对付取证收集的技术。数据隐藏包括隐秘术（Steganography）和数字水印 2 个主要研究方向。在反取证中应用的主要是隐秘术，它将秘密信息嵌入到看上去普通的信息中进行传送，以防止第三方的检测。也就是说，信息隐秘是通过建立传送秘密信息的秘密信道来实现的。可以对隐秘系统的隐秘性和鲁棒性这 2 个基本要求进行定量分析：

已知消息 M，载体 C 可看作一列随机变量，隐藏信息 K 通常假定为独立、均匀分布的随机变量，对给定的 K，由编码算法 f_k 产生的隐秘消息 $S = f_k(C, M)$ 仍为一随机变量，隐秘性要求 S 与消息源中的典型消息不可区分，这一条件可用编码约束 $Ed(C, S) < = \delta$ 来表示。其中 d 为失真度，E 是关于 C 和 S 的联合概率分布的期望算子。显然，界 δ 要足够小才能保证达到隐秘性。鲁棒性要求在各种扰动和攻击后，失真后的信息 S^1 仍然具有合理的质量，可用通道约束表示为 $Ed(S, S^1) < = \delta$。

对反取证而言，加密的缺点是直接表明通信是秘密的，而隐秘术的缺点是需要大量的开销来隐藏相对少的信息比特，即 K 和 S 的比值很小，效率很低，另外，一旦该系统被发现，就会变得完全没有价值。由此，先加密信息，再用隐秘术来隐藏，将使取证更加困难。文献[5]还指出了与新的加密技术相联系的更有效的使用 Slack 空间隐藏数据的方法。

3.4　数据转换技术

罪犯一旦攻陷了一个远程系统的组件，其主要目标就是尽可能长地保持对新的已获取资源的控制。对于管理和调查人员而言，在证实有未授权对象存在时会认为有网络入侵者侵入系统，这些对象包括文件、目录、进程、任务或网络连接等。为此，罪犯会采取数据转换（Data Transformation）技术将上述对象从常规的操作中隐藏，并保持或重建一个调查人员的信任，使调查人员迷失方向或产生错误的判别[6]。

3.5　数据混淆技术

混淆（Obfuscation）通常描述故意使某事物难以理解的行为。数据混淆则是一种逃避技术，主要指用特殊符号（如 Backspace、Insert 和 Delete 等控制符）来隐藏攻击。这种技术要求用十六进制形式表示字符以逃避检测。另外，使用 Unicode 表示法也是一个有效的逃避检测的方法。数据混淆使得罪犯能够在对程序和数据进行

访问的同时,限制调查人员对证据的识别和收集.

3.6 防止数据创建技术

防止数据创建(Data Creation Prevention)是一种有效的攻击取证收集的手段.在一个犯罪分子和信息系统的常规事务中,大量潜在的证据被建立.如果罪犯能够在调查人员识别、收集之前防止相关数据的创建,将增加他们成功避开被检测和取证的机会.直接的防止技术包括使用 Root Kits 或修改系统二进制(System Binaries)[7],但是,这种方法在限制数据创建的同时也限制了罪犯的攻击能力.

4 数据擦除反取证应用分析与设计

4.1 数据擦除技术

数据擦除(Data Wiping)用来定位用户磁存储介质上存储的活动记录,搜寻并不可逆地删除它们,是一种逻辑摧毁数据的方法.数据擦除起源于对用户隐私的保护,但目前已被犯罪分子用来进行反取证.

数据擦除的基本思想是用某种覆盖方案反复重写磁介质.典型的覆盖方案有3种:美国国家安全局推荐的7次随机覆盖方案[8];美国国防部 DOD 5220.22-M(C and E)标准推荐的3次0,1交替覆盖方案;Gutmann方案[2],它需要进行35次覆盖,被认为是最安全的擦除算法[9].前两者较简单,后者是针对当时的3种磁盘编码 MFM,PPM 和 PWM 提出的,数据擦除彻底,但缺点是速度太慢,且仅对 MFM 等3种编码效果较好.目前的一些数据擦除软件主要是基于简单的0,1交替覆盖方案,并不能完全清除所有数据[10].

4.2 基于 m 序列的数据擦除覆盖方案设计

现代编码技术可以用五元组$(d,k;m,n;r)$来表示不同的 RLL 码,其中,d、k 分别表示记录序列中相邻两个1之间至少有 d 个零、最多不得超过 k 个零,只要正确设计 d、k 值,即可获得优良的 RLL 编码性能.在近几年发展的高密度磁盘中,不同的 RLL 码得到了广泛应用.由于编码表示的统一,基于 Gutmann 方案的思想,笔者设计了基于 m 序列的通用覆盖方案.

m 序列是一种二进制伪随机序列,通常由线性反馈移位寄存器生成.设一线型反馈移位寄存器为 r 级,其序列多项式 $G(x)$可表示为

$$G(x) = a_0 + a_1x + a_2x^2 + \cdots + a_nx^n + \cdots = \sum_{n=0}^{\infty} a_nx^n \tag{1}$$

其中 a_n 是一个二元有限域{0,1}的元素,运算为模2和.显然,序列$\{a_n\}$与 $G(x)$一一对应.再根据 Caley-Hamiton 定理,可推出其特征多项式

$$f(x) = \sum_{i=0}^{r} c_ix^i \tag{2}$$

其中 c_i 为反馈系数,其取值也是一个二元有限域{0,1}的元素.通过变换,可推出序列多项式与特征多项式之间的关系为 $G(x) = x/f(x)$.由此可得到产生序列$\{a_n\}$的方法,即对 $f(x)$进行长除,得到序列多项式 $G(x)$,$G(x)$的系数就是所求序列.当该线性反馈移位寄存器的抽头系数取自一本原多项式时,所得序列即为一 m 序列.

根据 m 序列的性质,其周期 $L = 2^r - 1$,考虑到1个扇区的字节数为512,若 $L+1$ 对应于一个磁盘扇区字节数,则可得 r 为9,其本原多项式共有 $\Phi(2^9-1)/9 = 48$ 条,可从中选择反馈系数(1000010001),即得本原三项式

$$f(x) = x^9 + x^4 + 1 \tag{3}$$

它产生的 m 序列长度为511位,在最后加上与其前一位反转的值,就得到一个512位的序列 m_1.

m_1 可以用软件方法产生,根据 RLL(d,k)码的性质,在产生 m_1 时对2个1之间零的个数进行限制,即2个1之间至少有 d 个、最多不得超过 k 个零.这只需要在算法中增加0计数器 Count 0,当产生1个1时 Count 0开始对0计数,当下一个1产生时,若 Count 0 < d,则将该1反转,同时将 Count 0 加1,当下一个1产生前,若 Count 0 > k,则将该0反转,并重置 Count 0.由此可产生一个针对不同 RLL 码的以扇区为单位的磁翻转序列 m_2,而且,不同于流加密,这样产生的序列是不可逆的.一组 m_2^i($i = 1,\cdots,n$)即可组成一个覆盖方案,其覆盖次数 n 可根据用户对删除数据的重要性来确定.

显然,该方案有3个性质:(a)该方案是一种基于 RLL 码原理的通用覆盖方案;(b)覆盖是按扇区进行的;

(c)覆盖序列具有密码学性质.由于数据擦除的程度与覆盖次数及磁盘记录编码有关,因此本文所提出的方案,对于相同的覆盖次数而言,据性质(b)和(c),其安全性能和速度高于NSA方案;如果覆盖次数相同,据性质(a)和(b),其速度高于Gutmann方案且具有通用性.

5 结 束 语

目前的反取证技术还处于起步阶段,越来越多的安全技术与理论,尤其是为保护个人隐私所采取的加密、隐藏、数据擦除等技术,已经被应用于计算机犯罪和反取证.面对反取证技术的发展,计算机取证系统的设计需要进一步加强现有软件系统的防范措施.可以预见,取证与反取证的对抗必将越来越激烈,只有对反取证技术深入了解,才能不断提高取证软件的可靠性、可信任性.在某种意义上,保护隐私以外的反取证是一种犯罪行为,对反取证技术的研究将成为信息安全与取证技术研究的重要方向.

参考文献:

[1] JOHANSSON C. Forensic and anti-forensic computing[EB/OL]. 2002[2006-01-16]. http://www.fukt.bth.se/~uncle/papers.

[2] GUTMANN P. Secure deletion of data from magnetic and solid-state memory[C]//The Sixth USENIX Security Symposium. San Jose: [s.n.], 1996:23-35.

[3] FARMER D, VENERMA W. Forensic discovery[M]. Boston: Amazon, 2004: 146-147.

[4] CLEMENS F. TKS1—an anti-forensic, two level and iterated key setup scheme[EB/OL]. 2004[2006-01-16]. http://clemens.endorphin.org.

[5] RIPE P. Advanced anti forensics[EB/OL]. 2005[2006-01-16]. http://www.phrack.org/show.php.

[6] CHRISTIAN S J, MICHAEL L. Digital anti-forensics: emerging trends in data transformation techniques[EB/OL]. [2006-01-16]. http://www.seccuris.com/documents/papers.

[7] CALOYANNIDES J, MICHAEL A. Privacy protection and computer forensics[M]. Boston: Artech House, 2004:223-225.

[8] U S DoD. Standard 5220.22-M National industrial security program operating manual(NISPOM)[S]. Washington: GPO, 1995: 800-831.

[9] PETER B. Data security for surplus or end of life hardware[C]//2004 NYS Cyber Security Conference. New York: [s.n.], 2004:21-33.

[10] MATTHEW G. Evaluating commercial counter-forensic tools[C]//DWFS 2005. New Orleans: LA, 2005:106-118.

Research on computer anti-forensics

ZHANG You-dong[1,2], WANG Jian-dong[1], ZHU Wu-jia[1]

(1. *Department of Computer Science and Engineering, Nanjing University of Aeronautics and Astronautics, Nanjing* 210016, *China*;
2. *Department of Computer Engineering, Huaiyin Institute of Technology, Huaian* 223003, *China*)

Abstract: The concept and methodology about computer anti-forensics were analyzed. The main techniques of anti-forensics, such as data wiping, data hiding, data obfuscation, and data transformation, were compared. Then, a safe and high-efficient data wiping algorithm based on sequence m was proposed.

Key words: anti-forensics; data wiping; data hiding; data obfuscation; data transformation

Network Forensic Computing Based on ANN-PCA*

Youdong Zhang[1,2], Yongjun Ren[1]
[2] Department of Computer Engineering
Huaiyin Institute of Technology
Jiangsu Huaian 223001, China
z.yd@163.com

Jiangdong Wang[1], Liming Fang[1]
[1]College of Information Science and Technology
Nanjing University of Aeronautics and Astronautics
Jiangsu Nan Jing 210016, China

Abstract

The network forensic computing is faced with the question of the massive information stream analyses. Two solutions including feature extracting and classification mining were proposed. The main work includes features extracting with ANN-PCA, classification mining after features extracting in FAAR algorithm which we proposed to mine association rules based on PCA and Fuzzy classification technology. The experiment indicates that the classification accuracy raise distinctly after feature extracting, and the feature quantity decrease especially U2R attack. Further more, the number of generating rules also reduces clearly.

1. Introduction

1.1. Analysis of research methods

The electric evidence of network crime has the volatility character, and it has hided in massive information. So the problem how to automatic and effective find the electric evidence in massive data is urgently in network forensic computing domain. It is the same as abnormal detection of network intrusion area, the network forensic need proactive defense also. It is required to be able to detect the known network attack and forensic in real time, and also to identify the unknown attack or the abnormal behavior for forensic indention.

Generally, in order to analysis the electric evidence, the system needs to store all network information [1], but not all information is useful to the forensic analysis. If we were able to find the key feature, it will help to reduce the storage, increase the detection accurateness and speed. In statistic view, this is not to influence the veracity of result. There are two solutions may be considered:

One method is sample analysis. In the study of intrusion detection domain, researcher has proposed the sample analysis algorithm with statistic methods [2, 3], and referenced the data flow model in database domain to solve the network traffic monitor or abnormal detection for continuous, high speed and high dimension network traffic [4]. Although there are no report to design forensic system in this method, but it is a probably solution.

Another method is feature extraction, it is to say to extract classical character scheme from vast network information. There are two ideas in this method. Firstly, we consider the importance or correlation of the feature. Normally, intrusion data has high dimension, but the information about intrusion behavior is usually concentrated into partial attribution, for example, to DoS and Probe intrusion, they correlate with traffic attribution, and that U2R and R2L correlate with content attribution. If we used all attribute to detect intrusion, the classification accurateness will be affected by redundancy attribution, and the detection time will increase largely. Considering these factors, Sung proposal a ranking importance method for network forensic [5], he takes all the features of KDD CUP 99 dataset to import the SVM classifier, and ranks the significance of inputs with important, secondary or unimportant label, then carries forensics computing. On the other hand, Lee assigns some key features as constraint to mine the audit data [6]. Although there are no report to forensic computing in this method, but it may be a possible method.

Secondly, we can reduce the dimension of the features. Through transforming to the high dimension

* This paper is supported by the 863 important project, China (Grant Number 2006AA12A106), and the Academic Natural Science Research Project of Education Department of Jiangsu Province, China (Grant Number 06KJD520019)

data, we can obtain a new low dimension data. There are already some research results in intrusion detection domain. Our studies in this paper are enlightened by the work of Sung and Lee.

1.2. Related Work

Network forensic computing is the act of capturing, recording and analyzing network audits, network flow in order to discover the source of security breaches. Most of current techniques are passive monitors. In most of the cases new attacks signatures are detected manually or in some cases go undetected until the incident is being reported. The main focus in the area of network forensic computing is to automate the process of detecting all the attacks and to prevent the damages caused by further security breaches.

We propose a new network computing method based on ANN-PCA in this paper. The main idea of our work is to identify all possible security violations with ANN-PCA method, and build these signatures into the detection with association rule mining.

This paper is organized as follows: The following section analyses the PCA methods and propose a PCA computing algorithm based on ANN-PCA. In section 3 we analysis the experimental results, and we draw our conclusions in section 4.

2. Forensic feature selection based on ANN-PCA

2.1. Primary component analysis

Let raw data set D has n samples, $X=\{X_1, X_2, \quad, X_s\}$ is an n-dimension attribute variable set. According to the primary component analysis methods, we have:

$$Y_1 = w_{11}x_1 + w_{21}x_2 + \cdots + w_{s1}x_s$$

$$Y_2 = w_{12}x_1 + w_{22}x_2 + \cdots + w_{s2}x_s$$

.....

$$Y_s = w_{1s}x_1 + w_{2s}x_2 + \cdots + w_{ss}x_s$$

that is to say :

$Y=W^T X$ (1)

Here,

Var $(Y_i) = \lambda_i$, $i=1,2, \quad ,s$, $\lambda_1 \geq \lambda_2 \geq \quad \geq \lambda_s \geq 0$,

Cov$(Y_i, Y_j) = 0$, $i \neq j$, $i,j \leq s$.

Where λ and W are eigenvalues of PC's covariance matrix $\sum$ and its orthogonal eigenvectors respectively. Then, Y_1, Y_2, ... Y_s are called the first PC, the second PC, and the s PC of X respectively. In PCA, the contribution rate $\lambda_i / \sum_{j=1}^{s} \lambda_j$ denotes the i PC's proportion of extracting information from D, and the accumulative contribution rate $\sum_{i=1}^{m} \lambda_i / \sum_{j=1}^{s} \lambda_j$ denotes the ability for the first m PCs to extract information from the original data set. Commonly, the first m PCs accumulative contribution rate may have a high proportion (great than 85%), so we can use these m independent PCs rather than the original attributes for data analysis.

According prior definition, to solve the PCs has to compute the covariance matrix $\sum$ and its orthogonal eigenvectors. For the high dimension of intrusion data, the statistic method will spend huge CPU's time. That is not satisfied the real time forensic computing. So we propose a method to solve PCs with ANN-PCA technique effectively.

2.2. Learning the PCs based on ANN-PCA

Oja introduced the algorithm of extracting PCA with neural network in the first. But Oja's algorithm is only able to learn one PC. Cichocki proposed CRLS algorithm with neural network technique to extract several PCs for image identify application [7]. The neural network structure shows in Figure 1. It's an m level cascade neural network depending on the number of required PCs m. An appropriate number of m cascades can be stacked one after other. The learning principal component is performed in the same way as for the first component.

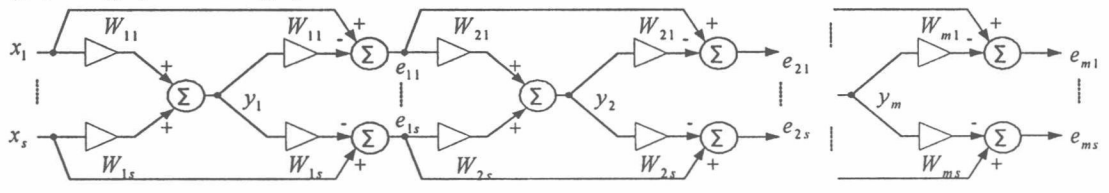

Figure 1. The architecture of a cascade network for PCA

In this structure, we assign the weight with orthogonal eigenvectors of formula (1) , e_i =(e_{i1}, e_{i2}, , e_{is}) is a loss function. It carry out the extraction process not directly from the input data x(t) but from the loss function computing. Then, we have iteratively to supply the input of this network with the training data until all weight vectors w_j (j=1, , m) are

stabilized. Through improving the epoch process of CRLS, the algorithm called ANN-PCA of learning m PCs is given in figure 2.

```
for k=1 to s
  e0(k)=x(k)          //initialization
for j=1 to m          //for every neuron
  {wj(0)=1
  ηj(0)= σ[ej-1]/s    //σ[e]=  ,  ke_tk^2
  for k=1 to s
    {yj(k)=wj(k-1)ej-1(k)
     ηj(k)= yj(k)^2+ ηj(k-1)
     wj(k)= wj(k-1)+[ yj(k)/ ηj(k)][ ej-1(k)- wj(k-1) yj(k)]
     if  | wj(k-1)- wj(k) |< ε
     { wj= wj(k)  ;exit}
        }
  }
  for k=1 to s
    yj(k)= wj ej-1(k)
    ej (k) =ej-1(k)- yj(k) wj
```

Figure 2. ANN-PCA algorithm

In the algorithm, $\eta_j(0)$ is initial learning rate and $\eta_j(k)$ is adaptively, $\sigma[e_j]$ is a individual loss function, and $\sigma[e]=\ ,\ k e_{tk}^2$.

3. Empirical results and comparisons

3.1 The data

The experiment use the famous KDD Cup 99 dataset [8], it is used to as the test data of the intrusion feature selection or classification study. The dataset has 41 features, including 34 quantitative and 7 qualitative features. Each sample is a TCP/IP connection record. Except normal pattern, the attack types fall into four main categories: Probe, DoS, R2L and U2R. According the general view, from this dataset a subset of 494021 data is used in our experiment, of which 20% represent normal pattern.

Some of data in connection record have large value range. For PCA, feature measure diverse makes the total variance be controlled by the features that have large value. It will lead to the irrationality results. So we normalize each feature value with the formula:

$$X_i=(X_i-\mu_i)/\sqrt{\sigma_{ii}}$$

3.2 Empirical results

The experiment has two steps. Firstly, all features are used to classification mining with FAAR algorithm directly; the results are shown in table 1. FAAR is an algorithm to mine association rules we propose in reference [9]. It can mine all kinds of rules including positive association rules, negative association rules and extra association rules we define based on PCA and fuzzy classification approach, and its PCAs are calculated with statistical methods.

Table 1. Result of full features testing

Class	Initial Features	Accuracy %	Rules
Probe	41	93.30	12
DoS	41	93.21	30
R2L	41	93.79	18
U2R	41	98.15	24

Secondly, we use ANN-PCA algorithm to extract features, and the extracting features are used to import to the FAAR algorithm for classification and mining. In FAAR algoritm, the threshold of accumulative contribution rate is 90%, the η cut set $\underset{\sim}{X}_{\eta}$ of $\underset{\sim}{X}$ is 0.5, the η cut set $\underset{\sim}{Y}_{\eta}$ of $\underset{\sim}{Y}$ is 0.5 also. The results under these conditions are shown in table 2.

Table 2. Result of feature extracting and classification mining

Class	Extracting Features	Accuracy %	Rules
Probe	28	98.26	7
DoS	28	98.13	18
R2L	33	98.96	10
U2R	20	98.78	14

Comparing table 2 with table 1, the classification accuracy raise distinctly after feature extracting, and the information quantity decrease especially for U2R attack. Further more, the number of generating rules reduces clearly. Though network forensic more focuses on evidence but the analysis efficiency and false positive rate [5, 10], the experiment indicates that false positive rate also decrease after feature extracting.

Currently, many researches of network forensic concentrate in forensic analysis to audit data, and almost belong to static analysis. Reference [10] designs a prefix system framework for online network forensic computing to audit logs based on Neyman-Pearson Lemma theory, but it hasn't experiment results. So we can't compare with it.

The work same as this paper is only references [5], [11] and [12]. Reference [12] proposes a network forensic model FENSF based on fuzzy expert system technology. It is only able to forensic computes to five kinds attack, TCP port SCAN, TCP SYN Flooding, ICMP smurf, Ping of Death and Land, and that all belong to DoS attack. Reference [11] designs a system

ADenoIdS to remote buffer overflow attack forensic, it belongs to R2L attack. Reference [5] is the most same as this paper; it ranks the features in SVM classifier with one of important, secondary or unimportant label. The number of features we find is more than its important, less then its important plus secondary, but mining rules use the important plus secondary features that the reference [5] suggested to apply in forensic computing, the generating rules is more than our work's obviously.

4. Conclusion

Network forensic computing is a crosses science research direction. It is faced with many questions such as massive information forensic, remote access to crime scenario, anti computer forensic and so on. We propose a feature extracting and classification mining method based on ANN-PCA to network forensic computing in this paper. The method improves the classification accuracy distinctly and decreases the information quantity especially for U2R attack. It will apply in the safety subsystem of network forensic in 863 important project which studies the scheduled flight delay alerting and affecting analysis. The next research includes increasing forensic computing and cooperation forensic computing.

5. References

[1]V. COREY, C. PETERMAN, and S. SHEARINS, "Network forensics analysis", IEEE Internet computing, 2002, 6(6), pp.60-66.

[2]P. DOMINGOS, and G. HULTEN, "Learning from infinite data in finite time", In T. G. Dietterich, S. Becker, and Z. Ghahramani, editors, Advances in Neural Information Processing Systems 14, MIT Press, Cambridge, MA, 2002, pp.673-680

[3]P. DOMINGOS, and G. HULTEN, "Learning from infinite data in finite time", In T. G. Dietterich, S. Becker, and Z. Ghahramani, editors, Advances in Neural Information Processing Systems 14, MIT Press, Cambridge, MA, 2002, pp.673-680

[4]B. BABCOCK, S. BABU, M. DATAR., R. MOTWANI, and J. WIDOM, "Models and issues in data streams", Proc. of the 21st ACM SIGACT-SIGMOD-SIGART Symp. on Principles of Databases Systems. Madison: ACM Press, 2002, pp.1-16

[5]S. MUKKAMALA, and H. A. SUNG, "Identifying significant features for network forensic analysis using artificial intelligent techniques", International Journal of Digital Evidence. Winter 2003, 1(4), pp.1-17

[6]W. LEE, S. STOLFO, and M. KUI, "A data mining framework for adaptive intrusion detection", In IEEE Computer Society,ed. Proceedings of the 1999 IEEE Symposium on Security and Privacy. Oakland, CA: IEEE Press, 1999, pp.120-132

[7]A. CICKOCKI., W. KASPRZAK, and W. SKARBEK, "Adaptive learning algoith for principal componet analysi s with partial data[C]", Trappl R (ed.),Cybernelics and Systems '96, volume 2. Austrian Society for Cybernetic Studues. Vienna, Austrua, 1996, pp.1014-1019

[8]S. HETTICH, and S.D. Bay, "KDD Cup 99 Task Description". http://kdd.ics. uci.edu/ databases /kddcup99/task.html,1999

[9]Z. YOUDONG, W. JIANDONG, C. HUIPIN, and Y. FEIYUE, "The multiple statistic analysis of association rule mining" [C].Proc. of AWFS'03, Nanjing, China, 2003, vol.1, pp.25-28

[10]P. G. BRADFORD, M. BROWN, J. PERDUE, and B. SELF, "Towards proactive computer-system forensics", In Information Technology: Coding and Computing, 2004. Proceedings. ITCC 2004. International Conference on 2004, (2), pp.648- 652

[11]H. JIAWEI, P. JAIN, and Y.YIWEN, "Mining frequent patterns without candidate generation: A frequent-pattern tree", Data Mining and Knowledge Discovery, 2004, 8, pp.53-87

[12]A. EVFIMIEVSKI, R. SRIKANT, and R. AGRAWAL, "Privacy preserving mining of association rules", Proc. of 8th ACM SIGKDD Intl. Conf. On Knowledge Discovery and Data Mining(KDD) , 2002, pp. 255-267